1-31-2013

A SCAFFOLED INSTRUCTIONAL MODEL TO TEACH DOCTOR OF PHARMACY STUDENTS TO EVALUATE RANDOMIZED AND NON-RANDOMIZED MEDICAL STUDIES ON A SIMILAR TOPIC TO REACH A CLINICAL CONCLUSION

Stefani Dawn

Follow this and additional works at: http://digitalrepository.unm.edu/educ_llss_etds

Recommended Citation
Dawn, Stefani. 'A SCAFFOLED INSTRUCTIONAL MODEL TO TEACH DOCTOR OF PHARMACY STUDENTS TO EVALUATE RANDOMIZED AND NON-RANDOMIZED MEDICAL STUDIES ON A SIMILAR TOPIC TO REACH A CLINICAL CONCLUSION.' (2013). http://digitalrepository.unm.edu/educ_llss_etds/53

This Dissertation is brought to you for free and open access by the Education ETDs at UNM Digital Repository. It has been accepted for inclusion in Language, Literacy, and Sociocultural Studies ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact disc@unm.edu.
This dissertation is approved, and it is acceptable in quality and form for publication:

*Approved by the Dissertation Committee:*

Rosalita Mitchell, Chairperson

William Troutman

Diane Velasquez Torres

Richard Bryant
A SCAFFOLDED INSTRUCTIONAL MODEL TO TEACH DOCTOR OF PHARMACY STUDENTS TO EVALUATE RANDOMIZED AND NON-RANDOMIZED MEDICAL STUDIES ON A SIMILAR TOPIC TO REACH A CLINICAL CONCLUSION

BY

STEFANI DAWN

B.S., Anthropology, University of Texas at Austin, 1994
M.A., Education: Language, Reading and Culture, 1995
M.S., Soil, Water and Environmental Science, 1998

DISSERTATION

Submitted in Partial Fulfillment of the Requirements of the Degree
Doctor of Philosophy

Multicultural Teacher and Childhood Education

The University of New Mexico
Albuquerque, NM

December, 2012
ACKNOWLEDGEMENTS

Twenty-two years ago, in 1990, I heard someone say that you do not really pursue a PhD, rather, that it pursues you. I do not know why those words stuck with me. At the time I did not understand them, nor could I relate to them. Had I been asked then if I would get a PhD, my response would have been, “no way.” At the time, I was mired in getting a Bachelor’s Degree. That was enough. Even years later, after two master’s degrees, I actively said “no” to a PhD. But, little did I know, somewhere in the recesses of my mind and spirit, I was being pursued. Now, with the years of PhD-related work, tears, and feelings of wanting to quit behind me, I get it. I was chased down until I gave in and signed on the dotted PhD-line. I could only keep moving forward until the dissertation was done. Indeed, I was pursued.

Such a monumental effort is not done in a vacuum. It is done with the guidance, support, and patience of many people. These people deserve my acknowledgement and heart-felt gratitude. In the professional realm I thank my committee members, Drs. Rose Mitchell (chair), Diane Velasquez Torres, Richard Bryant, and William Troutman. Of special note, Dr. Mitchell helped pick up the ball when my momentum was the lowest. Dr. Troutman was a true mentor. He has worked with me since the earliest days of the pilot study for this work and guided my thinking around evidence-based medicine. He also modeled tact, perseverance, and humor in the presence of the unplanned challenges that life presents.

I also acknowledge the additional College of Education faculty who were part of my comprehensive exam, proposal hearing, and the pharmacy faculty who participated in
the pilot research, and this dissertation research. All of you propelled me along the path, thank you.

In the personal realm there are many people I must acknowledge. Those who are closest to me deserve a thank you that goes far beyond these pages and words, to and from my very being and the future that rests with it. You were with me in the daily grind, experiencing and supporting me in the most challenging of moments - not just for hours or days, but for years. You deserve my deepest, most heart-felt thank you and love - Dean, Bob, Greggor, Dad, and Mom.

There are also those who were lights along the path, providing energy shots and motivational pushes. You helped keep up the momentum, each holding special memories and contributions. Thank you to: my very dear friend and co-PhD torch-bearer, Michelle; my newest friend and neighbor, Ken, who is a model for thoughtfulness; “Hawaii Bill,” who gave me the best “you can’t quit” lecture via email ever, if you only knew how much that impacted me (perhaps you do now); and my best friends, Pat, Julie-Ann, and Brian—you all are foundational.

Thank you to all, thank you.
A SCAFFOLED INSTRUCTIONAL MODEL TO TEACH DOCTOR OF PHARMACY STUDENTS TO EVALUATE RANDOMIZED AND NON-RANDOMIZED MEDICAL STUDIES ON A SIMILAR TOPIC TO REACH A CLINICAL CONCLUSION

By

Stefani Dawn

B.S., ANTHROPOLOGY, UNIVERSITY OF TEXAS AT AUSTIN, 1994
M.A., EDUCATION: LANGUAGE, READING AND CULTURE, 1995
M.S., SOIL, WATER AND ENVIRONMENTAL SCIENCE, 1998
PHD, MULTICULTURAL TEACHER AND CHILDHOOD EDUCATION, 2012

ABSTRACT

The goal of this study was to test a scaffolded instructional model in a complex Evidence Based Medicine (EBM) lesson where students evaluated three different medical studies (RCT, case-control, and cohort) on the same topic and used those studies to reach an evidence-based conclusion. The hypothesis was that the students’ literature evaluation skills and subsequent application of the literature to address a clinical question would more closely approximate the experts following implementation of the model. The results do not fully support the hypothesis. The conclusions were: (1) third-year doctor of pharmacy students at the college have a limited ability to evaluate medical literature of varying qualities and types and conflicting conclusions; (2) prior to reading the RCT, students’ initial clinical conclusions more closely resembled the experts’, potentially indicating an unbalanced influence of the RCT, either from RCT bias/preconceptions or a lack of skills transfer in evaluating the RCT; and (3) the instructional model needs further development by adding explicit instructional scaffolding around the Medical Literature Evaluation (MLE) Rubric, vocabulary, and directly addressing student preconceptions/biases.
# TABLE OF CONTENTS

## CHAPTER 1 INTRODUCTION

- Background of the Problem
- Evidence Based Medicine (EBM)
- History of EBM in Instruction
- Presence of EBM at a Public College of Pharmacy (COP)
- Significance of the Study
- Study Objectives and Hypothesis
- Overview of the Dissertation Chapters
- Definitions

## CHAPTER 2 REVIEW OF RELATED LITERATURE

- Introduction
- Educational Theory: Constructivism, Adult Learning Theory, and Science Literacy Theory
- Constructivism
- Adult Learning Theory
- Science Literacy Theory and Information Literacy Theory
- Instructional Approaches: Scaffolding, Science Instruction and Reading Scientific Text
- Publications Describing the Incorporation of Primary Scientific Literature in Instruction
- Typical Publications in Teaching Using Primary Literature/EBM
Journal Clubs .............................................................................................................36

Publications in Teaching Using Primary Literature/EBM that Consider Education Research and Theory .................................................................38

Evidence Based Medicine Literature ..................................................................43

Literature to Inform the Development of the Medical Literature Evaluation (MLE) Rubric .....................................................................................43

Synthesizing Research to Reach a Clinical Conclusion ..................................50

CHAPTER 3 METHODS ...............................................................................................58

Study Background ....................................................................................................58

Study Objectives ......................................................................................................62

Study Population ......................................................................................................63

Major Eligibility Criteria .........................................................................................64

Research Procedures Involving Participants .........................................................65

Study Procedures .....................................................................................................66

Risks and Discomforts ..............................................................................................71

Confidentiality and Privacy .....................................................................................71

Methods for Obtaining Expert MLE Scores, Article Rankings, and Clinical Conclusions .....................................................................................72

Selection of the Medical Articles ............................................................................72

CHAPTER 4 RESULTS ...............................................................................................75

Consents ...................................................................................................................75

Characterization Survey ..........................................................................................75
MLE Rubric Scores, Article Ranking, Crux Vocabulary, and Initial Clinical

Conclusion.................................................................82

MLE Rubric Scores......................................................84
MLE Rubric Matching Score Analysis.................................95
Article Ranking..........................................................102
Crux Vocabulary..........................................................104
Clinical Conclusion .....................................................110

Scaffold Contributions Survey........................................114

Limitations of the Study Design......................................118

CHAPTER 5 DISCUSSION..................................................120

Conclusion One..............................................................124

Student Challenges with Analyzing the Medical Literature – The MLE
Rubric.................................................................124

Student Challenges with the Crux Vocabulary.........................126

Conclusion Two..............................................................128

Student Ability to Reach a Clinical Conclusion.........................128

Conclusion Three...........................................................130

The MLE Rubric...........................................................131

The Crux Vocabulary.......................................................133

The Small and Large Group Discussions..............................137

Implications for the Field.................................................144

Future Research...........................................................145

Conclusion.................................................................145
Chapter 1

Introduction

Evidence based medicine (EBM) is the practice of utilizing current, high quality scientific and best evidence in individual patient care, be it in medicine, nursing or pharmacy. Higher education accreditation agencies recognize the importance of EBM in the quality of medical delivery and are now requiring its incorporation into the professional curriculum. Although there are publications in professional medical-related education describing EBM instructional activities, much of it is not grounded in educational theory, learning theory, or evidence-based instructional practices. Many of those publications are also weak in providing evidence of student learning.

The goal of this study was to test a scaffolded instructional model in a complex EBM lesson where students evaluated three different medical studies (RCT, case-control, and cohort) on the same topic and used those studies to reach an evidence-based conclusion. The hypothesis was that the students’ literature evaluation skills and subsequent application of the literature to address a clinical question would more closely approximate the experts following implementation of the model. An outcome of this work seeks to improve college-level instruction in EBM such that students are demonstrating competency in medical literature evaluation and critical application of study conclusions to application of a clinical issue.

The instructional model for this work is based upon best instructional practices in science education (scaffolding, text coherence, vocabulary, and metacognition) and adult learning and incorporates important EBM concepts. The theoretical framework is grounded in constructivist, adult learning, and science and information literacy theories.
Figure 1 shows the foundational elements of the model, each of which is explained in more detail in subsequent chapters.

Figure 1. Foundational Elements for the Model

- **Evidence Based Medicine (EBM) - Goal**
  - Critical appraisal of the literature
  - Interpreting and applying evidence to the clinical issue.

- **Instructional Models**
  - Scaffolded (targeted incorporation of individual, small group, large group, and instructor mediated activities around science reading comprehension)
  - Reading comprehension for science text (text coherence, vocabulary, metacognition)

- **Educational Theories**
  - Constructivist learning theory
  - Adult learning theory
  - Science and information literacy theory

The initial version of the instructional model (Figure 2) was tested in a pilot study in 2007. Students were asked to first critically evaluate one type of medical study, an RCT, following traditional lecture-based didactic instruction. Then following implementation of the Clinical Literature Evaluation activity and instructional model, the students were asked to re-evaluate the study.
The pilot showed statistically significant improvements in Doctor of Pharmacy students’ ability to more accurately evaluate an RCT following the implementation of the scaffolded instructional model compared to their abilities following the didactic course alone (Dawn, Dominguez, Troutman, Bond, & Cone, 2011).

This dissertation research extends that original work into the more complex but necessary area of teaching students how to evaluate multiple types of studies on a shared topic to reach an accurate, clinically responsible conclusion. This work resulted in new instructional materials and was tested in a third year Doctor of Pharmacy course. The quality and accuracy of the student’s reviews of specific medical literature and the
accuracy of clinical conclusion they reach from that literature was investigated using qualitative and quantitative research methods. The study addressed the primary question:

- How do third-year Doctor of Pharmacy students’ ability to accurately evaluate three types of medical literature (RCT, case-control and cohort study) and reach a clinical conclusion compare to experts after scaffolded elements of the activity are implemented?

**Background of the Problem**

**Evidence-Based Medicine (EBM).** EBM is the practice of utilizing current, high quality scientific and best evidence in individual patient care, be it in medicine, nursing or pharmacy and is necessary for today’s health care practitioners to provide a high quality of patient care. There are five core EBM components, each of which requires its own skills, training, and definitions regarding what constitute best practice. The EBM components are: (a) accurately assessing a patient; (b) asking appropriate clinical questions to implement a meaningful literature search; (c) acquiring a body of high quality literature; (d) appraising the literature on validity, importance and usefulness (which is the focus of this dissertation), and; (e) interpreting and applying the evidence to the clinical issue (also the focus of this dissertation) (retrieved from JAMAevidence.com).

Professional medical organizations develop the details and expectations for EBM practice. In recent years, several national and international organizations have dedicated centers and resources for EBM research and training, such as the Centre for Evidence Based Medicine at the University of Oxford, the American Medical Association’s “JAMAevidence,” and the Cochrane Collaboration, which publishes the Cochrane
Library of medical reviews. These resources and EBM-based publications have expanded, refined and defined EBM including, how practitioners can be effectively trained and efficiently implement EBM, and how health care students should be trained to usher in a new era of evidence-based practice.

**History of EBM in instruction.** Although EBM began in the 1970s as an outcome of the increased availability of medical literature, it did not become a widespread movement in medicine until the early 1990s (Claridge & Fabian, 2005). As EBM adoption in medical practice grew, accreditation agencies for medical, nursing and pharmacy education began to require EBM instruction in the curriculum (Accreditation Council for Pharmacy Education, 2006; Wyer, et al., 2004). Despite the broad awareness of EBM and the conceptual support of it in the medical fields, the extent of formal EBM instruction in medical education has been limited.

A national survey of medical schools (n=269) conducted in 1998 showed that only approximately 37% of the internal medicine residency training programs in the United States (U.S.) provided a freestanding EBM curriculum (Green, 2000) (an update to the study has not been conducted). According to another study, U.S. emergency medicine residency training programs provided an average of 5 hours of EBM training per year (Kuhn, Wyer, Cordell, & Rowe, 2005), half the amount experts believe is needed to develop basic EBM skills (Kuhn, et al., 2005).

The 1998 survey by Green (2000) revealed that much of the EBM training is placed in “uncommitted” parts of the medical curriculum via small insertions into different courses or done through informal journal clubs outside of class. It also found that, although there are many medical education publications about implementing
successful journal clubs, clubs typically occurred outside of the formal curriculum and varied in rigor often resulting in a superficial introduction to EBM (Green, 2000).

Dirschl, Tornetta, and Bhandari (2003) provided a review of medical journal-club related articles and summarized the findings reaching similar conclusions as Green. They stated that “although most training programs sponsor journal clubs, they are not equally successful in meeting educational goals” and that “designing a format to make journal clubs stimulating, interesting, and educational for its members has proven to be a greater challenge than residency programs might have imagined” (p. 147).

Dirschl, et al. described “key considerations in the design, conductance, and evaluation of a journal club” (p. 147) including having leadership for the club in its organization, defining explicit goals (e.g. teach critical appraisal skills), having faculty experts and statisticians present (which appears to improve attendance), linking the journal club to the curriculum, and using structured review instruments to “guide the resident through the critical appraisal of the journal club article” (p. 151).

Even in schools with dedicated EBM courses, many curricula lacked “important structural elements” necessary for high quality EBM education (Green, 2000). For example, many of the EBM courses reviewed utilized only one or fewer of the top resources for systematic reviews (e.g. Best Evidence and the Cochrane Library) and had a shortfall in the formal development of the five areas of EBM (Green, 2000). Another shortfall was that many of the EBM curricula in medical schools had little or no internal evaluation to determine whether their educational objectives have been met (Green, 2000).
No studies were found characterizing the prevalence or quality of EBM instruction in pharmacy curricula.

**Presence of EBM at a Public College of Pharmacy (COP).** The pilot study for this work and this dissertation-specific research occurred at a COP in a large public University in the southwestern United States. The COP is accredited by ACPE and is, therefore, required to meet those accreditation standards. As part of those standards, every college of pharmacy in the United States must have clearly defined and measurable competencies that delineate the expectations for knowledge, skills and values they must posses upon graduation. Thus, the COP has 30 competencies, five of which refer to EBM:

1. Integrate and utilize knowledge of biochemistry, physiology, pathophysiology, and anatomy in order to design a pharmaceutical care plan. Acquire, comprehend, synthesize, apply, and evaluate information about the chemical structure and pharmacology of therapeutic agents in order to design, implement, monitor, evaluate, and adjust pharmaceutical care plans that are patient specific and evidence based.

16. Develop population-specific, evidence-based, and effective disease prevention and management programs.

17. Develop and implement population-specific and evidence-based disease management programs and protocols based upon analysis of epidemiologic and pharmacoeconomic data, medication use criteria, medication use review, and risk reduction strategies.

22. Evaluate the biomedical literature with regard to the
pharmacokinetics and pharmacodynamics of drugs.

27. Retrieve, analyze, and interpret the professional and lay literature to provide drug information to patients, their families, as well as other healthcare providers and the public.

The COP takes its role in providing the best education for its students seriously. This means that it strives to teach its competencies thoroughly, in a stage-appropriate manner and using best instructional practices. Prior the pilot study for this research the COP taught EBM in a traditional lecture setting and it taught the formal evaluation of medical literature to a limited degree. As a result of the pilot, the COP modified and expanded its instruction in EBM.

Significance of the Study

In addition to modifying and expanding EBM instruction in the COP Doctor of Pharmacy curriculum, this research provided additional data on the effectiveness of the scaffolded instructional model from the pilot study and identified changes that need to be made to the model to address the complexity of evaluating three types of studies on the same topic. It also revealed the skills students’ still lacked in critically evaluating the medical literature, as well as the potential influence of bias or preconceptions on students’ clinical conclusions. Although not a statistically significant change, initially 47% of the students chose the more conservative clinical treatment based upon the results of the cohort and case-control studies, and then, despite the presence of notable flaws in the RCT, that number dropped to 33% after students read the RCT. A statistically significant change in the student clinical conclusion that occurred after reading the RCT was a shift away from multiple/combination treatment options (43%), which was more
similar to the experts’ clinical conclusion, to singular treatment options (23%). This shift potentially demonstrates a lack of skills transfer in the students to critically evaluate the RCT and/or too much influence of the RCT compared to the cohort and case control studies (RCT bias).

It is anticipated this work will meaningfully contribute to the medical education field introducing evidence-based best instructional practices and research methodology that measures learning. It will also provide EBM-specific activities for instructors and substantive literature evaluation rubrics that are designed for students.

The pilot work has already been embraced by the larger pharmacy education community having been published in the peer-reviewed American Journal of Pharmacy Education (Dawn et al., 2011). The scaffolded instructional approach and associated medical literature evaluation rubrics are useful for pharmacy education specifically and medical education, including physician and nursing, more broadly. The scaffolded instructional approach and the associated concepts with the literature evaluation rubrics could also be used stepping-stones in advancing graduate-level education in a variety of research-based fields, including educational research.

**Study Objectives and Hypothesis**

The goal of this study was to test a scaffolded instructional model in a complex Evidence Based Medicine (EBM) lesson where students evaluated three different medical studies (RCT, case-control, and cohort) on the same topic and used those studies to reach an evidence-based conclusion. The hypothesis was that the students’ literature evaluation skills and subsequent application of the literature to address a clinical question would more closely approximate the experts following implementation of the model.
The study objectives were to:

- create, implement and edit a new Medical Literature Evaluation (MLE) Rubric to address case-control, cohort and randomized control trial studies.
- create, implement and edit a new lesson teaching students how to evaluate three different studies and types of studies (RCT, cohort, and case control) on the same topic, extract relevant data and reach an evidence-based clinical conclusion.
- observe implementation of the lesson in the third year lab to note any differences in delivery between the sections and to note potential modifications to improve lesson delivery;
- develop and implement a characterization-survey (based upon a survey implemented in 2007 for the pilot study) measuring the perceptions of student confidence in reading, evaluating, understanding and applying results from multiple medical studies towards a clinical conclusion;
- develop and implement a post-activity survey asking students to rate the different instructional elements/scaffolds in the activity; and
- evaluate students’ clinical conclusions by comparing their conclusions with expert clinical conclusions.

**Overview of the Dissertation Chapters**

The contents of this dissertation are presented in the following order.

- The end of Chapter 1 provides definitions of terms used in this document.
- Chapter 2 comprises the Literature Review which presents key educational theories upon which this work is based (constructivist, adult learning, and scientific and information literacy theories) and analyzes and synthesizes the
literature and gaps in the literature related to: (1) best instructional practices in science education (scaffolding and reading science text); (2) publications describing college and high school level instruction that used or incorporated scientific or medical literature; and (3) applying EBM themes related to evaluating and synthesizing medical literature to the instructional model tested for this dissertation research.

- Chapter 3 provides a brief description of the model, pilot study research and results and the methodology used for this research including the study design, description of the setting, data collection instruments and procedures, human subjects/consent, and statistical/data analysis procedures.

- Chapter 4 presents the results of the study (characterization survey, MLE Rubric scores, article ranking, crux vocabulary, clinical conclusion, and scaffold contributions survey).

- Chapter 5 provides a discussion of the results and changes that will be made to the instructional model as a result of the discoveries in the study.

Definitions

**Accreditation Council for Pharmacy Education (ACPE)** - the national agency for the accreditation of professional degree programs in pharmacy and providers of continuing pharmacy education (retrieved from [http://www.acpe-accredit.org/about/default.asp](http://www.acpe-accredit.org/about/default.asp))

**case-control study** - a study where patients with a disease or condition are identified and matched with a control group who do not have the disease. Data are collected retrospectively via medical records and interviews (Greenhalgh 2006).
**Case reports/case series** – case reports “describe the medical history of a single patient in the form of a story” (Greenhalgh 2006, p.52). A case series is when multiple case reports on a specific aspect of the same disease or condition are assembled to create a more comprehensive picture. Case reports are lowest in the traditional hierarchy of evidence.

**Clinical literature** (synonyms: clinical research publication) – refers to a randomized controlled clinical trial published in a medical or scientific journal.

**Clinical Literature Evaluation Rubric** (CLE Rubric) – a rubric developed by Dawn, et al. (2011) to help guide students through the process of evaluating the quality of a randomized control trial (RCT).

**Cohort study** – two or more groups are selected and observed related to the development a specific disease or outcome. One group is usually selected based upon their risk level or exposure to a particular agent (e.g. medicine or environmental exposure) and the other group serves as a comparison group who has not been exposed or lacks specific risk factors (Greenhalgh, 2006).

**Consolidated Standards for Reporting Trials (CONSORT)** – “an evidence-based, minimum set of recommendations for reporting” randomized control trials (RCTs) to aid in “complete and transparent reporting” which aids in the critical appraisal and interpretation of that work (Moher et al., 2001, 2010).

**Evidence-based medicine (EBM)** (synonym: evidence-based practice) - EBM is the practice of utilizing current, high quality scientific and best evidence in individual patient care, be it in medicine, nursing or pharmacy. There are essentially 5 core EBM components: (1) accurately assessing a patient; (2) asking appropriate clinical questions
to implement a meaningful literature search; (3) acquiring a body of high quality
literature; (4) appraising the literature on validity, importance and usefulness (which is
the focus of this dissertation), and; (5) interpreting and applying the evidence to the
clinical issue (retrieved from JAMAevidence.com).

hierarchy of evidence – refers to the relative weight that is given in the
biomedical field to the different types of medical research when trying to make evidence-
based medical decisions. In general, more weight is given to rigorous studies that
minimize the risk of bias (i.e. use random sampling techniques, have large sample sizes,
and have control groups). In general the highest weight is given in the following order:
systematic reviews/meta-analysis, randomized control trials (clinical trials), cohort
studies, case-control studies, cross-sectional surveys, case series, individual case reports
(Greenhalgh, 2006).

information literacy - Information literacy is the set of skills needed to find,
retrieve, analyze, and use information (retrieved from
http://www.ala.org/ala/mgrps/divs/acrl/issues/infolit/overview/intro/index.cfm)

medical literature – a broad term that refers to a publication of any type of
medical study (systematic reviews/meta-analysis, randomized control trials/clinical trials,
cohort studies, case-control studies, cross-sectional surveys, case series, individual case
reports)

Medical Literature Evaluation Rubric (MLE Rubric) – A rubric developed by
Stefani Dawn based upon the original CLE Rubric (Dawn, et al. 2011) that guides
students in evaluation three different types of medical literature research, the RCT, case-
control, and cohort studies. This rubric is based heavily on the CONSORT.
**meta analysis** – a statistical process that combines the quantitative results of multiple published studies on the same topic to increase the power and confidence in testing a specific hypothesis.

**scaffolded instruction** - "the systematic sequencing of prompted content, materials, tasks, and teacher and peer support to optimize” independent learning and can include instructional elements such as guided questioning, comparing ideas, identifying connections and distinguishing characteristics between concepts, and identifying valid relationships (Davis and Linn, 2000)

**science literacy** - the knowledge and understanding of scientific concepts and processes required for personal decision making, participation in civic and cultural affairs, and economic productivity (retrieved from [http://www.nap.edu/openbook.php?isbn=0309053269&page=22](http://www.nap.edu/openbook.php?isbn=0309053269&page=22)).

**scientific literature** – refers to an article published in a peer reviewed scientific journal.

**systematic review** – a type of publication that provides an update on the state of the science on a specific topic by assembling articles that topic, screening them for quality or specific study features (e.g. a randomized study) for inclusion in the review, and summarizing the results.
Chapter 2

Review of Related Literature

Introduction

In order to develop an evidence-based best-practices instructional model to teach Doctor of Pharmacy students how to evaluate medical literature and reach a sound clinical conclusion from multiple studies of differing types, there are several educational areas that require investigation. The educational theories that form the basis for this work are constructivism, adult learning theory, and science and information literacy theory. From these theories come the specific instructional practices (scaffolded instruction, science reading principles, and adult learning instruction) that guide the model to deliver key information in EBM (Figure 3).

Figure 3. Instructional Model Foundations

This literature review presents key publications in each of these areas and provides summaries of the conclusions and how these conclusions were used to guide the model. It also identifies weaknesses and gaps in the literature that may be addressed by this research.
The information is presented in the following sections:

- Educational Theory: Constructivism, Adult Learning Theory, and Science Literacy Theory
- Research on Specific Instructional Approaches: Scaffolding and Reading Scientific Text
- Review of Publications Describing the Incorporation of Primary Scientific Literature in Instruction
- EBM Literature to Inform Activity Development

**Educational Theory: Constructivism, Adult Learning Theory, and Science Literacy Theory**

*Constructivism.* Research on learning and cognition in science education tends to be based in constructivist learning theory, which posits that students construct knowledge in light of their exiting knowledge via the processes of generation, integration, and transformation of their experiential world (Gao, Baylor, & Shen, 2005).

Constructivism’s origin is often credited to Jean Piaget, but John Dewey, Maria Montessori, and Lev Vygotsky were also significant contributors to this thinking.

Piaget introduced the concepts of “assimilation” and “accommodation” proposing that when assimilation occurs in a learner new experiences are incorporated into existing frameworks. When accommodation occurs, which is often associated with failure, new frameworks are created or existing frameworks modified. More recent educational research specifically targeting science learning, builds from Piaget’s framework concepts showing that in order for students to develop competence in scientific study they must “(a) have a deep foundation of factual knowledge, (b) understand facts and the ideas in
the context of a conceptual framework, and (c) organize knowledge in ways that facilitate retrieval and application” (National Research Council, 2006).

The movement from constructivist learning theory to its application in instruction has taken numerous paths, many of which are lumped into the broad category of “active learning.” But what constitutes effective instructional techniques in “active learning” is being clarified through research. For example the application of constructivist learning theory in science instruction has fallen into two general approaches, inquiry-based instruction and scaffolded instruction – both “active learning.” Inquiry-based instruction, in its purest form, is entirely student-and-question-driven, while scaffolded instruction, is highly structured by the teacher with strategically-based supports and prompts for students through out the learning process (Dickson, Chard, & Simmons, 1993).

Inquiry-based approaches are frequently advocated for science instruction because they mimic the inquiry-driven nature of scientific research. Scientific research is a process that blossoms from a research question, which then generates subsequent questions and investigations as the scientist tries to unravel the mystery. Inquiry-based instruction has a similar open-ended approach as scientific inquiry.

Inquiry certainly has its place in science instruction, particularly when developing science research/experimentation skills, but educational research is showing that it is not as effective as instructional scaffolding in certain circumstances, particularly those that are heavily knowledge-dependent or complex and involve multi-layered learning which is applicable to this work (Chen & Bradshaw, 2007; Linn, Clark & Slotta, 2003; Collins, Brown, & Newman, 1989; Davis & Linn, 2000).
Instructional scaffolding is based upon Lev Vygotsky’s “zone of proximal development” (ZPD) which refers to the distance between the stage of independent learning and the developmental stage where instruction is needed (Vygotsky, 1934/1987).

According to Chaiklin’s analysis of Vygotsky’s ZPD:

The common conception of the zone of proximal development presupposes an interaction between a more competent person and a less competent person on a task, such that the less competent person becomes independently proficient at what was initially a jointly-accomplished task (2003, p.2).

This interaction between the more experienced (i.e. teacher) and less experienced (i.e. student) is where much of the instructional literature and research in scaffolding is focused, typically addressing the question “what instructional approaches are most effective?”

Scaffolding can include a variety of instructional elements such as guided questioning, comparing ideas, identifying connections and distinguishing characteristics between concepts, and identifying valid relationship all towards the goal of developing an independent learner (Davis & Linn, 2000). Scaffolds appear to be effective because they are explicit in their instruction and can help students develop conceptual frameworks that facilitate retrieval and application during the delivery of information.

Linn and colleagues developed and tested a scaffolded instructional framework for science. They found that four elements are important to have in instruction (Davis and Linn, 2000, p.820):
1. Making thinking ‘visible’ to students by illustrating how links and connections are made;

2. Making science accessible and identifying models of scientific phenomena that make sense to students so they can connect new information to existing knowledge and to problems that are both familiar and relevant;

3. Providing social supports so all students learn new links and connections for their ideas from their peers; and

4. Encouraging students to become autonomous learners so they can regularly visit their ideas and continue engaging in knowledge integration.

These features are considered in the instructional model associated with this dissertation. This is done in several ways, the instructor guided elements of the lesson make thinking visible by sharing the reasoning for specific decisions around scoring an article a certain way. Making science accessible to the students is done by using a topic with which they are familiar (cardiac pharmacology) and connecting it to the learning that occurred in their second-year coursework around medical research and literature evaluation. Social supports are provided through small group activities and the MLE Rubric is intended to help students become autonomous learners.

**Adult learning theory.** Since this work is being conducted in a Doctor of Pharmacy program, the instructional design needs to also consider principles in adult learning theory. The concept that adults differ from children as learners gained its footing early in the 1970’s and is largely credited to Malcolm Knowles from his book *The Adult Learner: A Neglected Species* (1973). In his work Knowles argued that adults operate
differently as learners from children. He cites research on human development stating that:

as an individual matures his *need* and *capacity* to be self-directing, to utilize his experience in learning, to identify his own readiness to learn, and to organize his learning around life problems, increases steadily from infancy to pre-adolescence, and then increasingly rapidly during adolescence (p. 43).

Knowles went on to state that the assumptions about learners made in pedagogy, or the traditional teaching of the day primarily focused on children and youth, are not appropriate for the adult. Thus, adult instruction needs to operate under different assumptions, which he calls andragogy. Knowles formally presented four assumptions underlying his androgogical theory:

1. Adults are self-directed. Thus, if an adult is not allowed to be self-directed in his learning experience, tension and resistance develops.

2. Adults have a plethora of experiences that provide a broad base to which new learning can connect. This experience also forms the basis of an adult’s identity. Knowles states that “to a child, experience is something that happens to him, to an adult his experience is *who he is*” (p. 46). Thus, if an adult’s experience is devalued or ignored, it is an assault on the person.

3. Adults learn what they *need* to because of their roles in society, like work, politics, parenting, etc., as opposed to children who learn what they “ought” to because of their biological and academic development.
4. Most adults enter education because they need to fill a gap related to a current situation and there is a desire to immediately apply what is learned. Knowles calls this a “problem-centered” orientation, compared to a “subject-centered” orientation that occurs in pedagogy, where the practical application of what is learned is postponed (e.g. children learn a subject to get into high school at some point into the future).

Knowles proposed instructional approaches that may better meet the assumptions about adult learners described above. He credited Tough’s work on project-oriented instructional approaches for adult learning as a major influence (as cited in Knowles, 1973). He also synthesized Hilgard’s 20 different instructional principles derived from three different families of theories, S-R theory, cognitive theory, and motivation and personality theory (as cited in Knowles, 1973), and other research around adult learning such as Rogers’ book *Freedom to Learn* (as cited in Knowles, 1973). In his work, Rogers presented the concept of the teacher as a facilitator of learning, which involves “realness,” respect of the learner, empathetic understanding, and accurate listening on the part of the teacher (as cited in Knowles, 1973). Knowles also referenced Houle, Dewey, and Bruner, among others, as having a strong influence on his work (as cited in Knowles, 1973). Knowles defined the elements of andragogy that meet the needs of the assumptions about adult learners (1973). He stated that:

- the climate necessary for andragogy is one that is informal and operates out of mutuality, respect, and collaboration;
- instructional planning by the teacher needs to incorporate the student, there needs to be a “mechanism for mutual planning” (p. 104);
the diagnosis of a learner’s needs should consist of “mutual self-diagnosis” (p. 104);

- the instructional design needs to be sequenced in terms of readiness (as opposed to subject matter) and it needs to be problem-centered (as opposed to content-based units);
- activities needs to be experiential (rather than transmitted); and
- evaluation needs to be mutual, “mutual measurement” and “mutual re-diagnosis.”

Since Knowles’ landmark publication, work in adult learning has taken many different directions including, exploring substantive shifts that can occur in how an individual sees him or herself as a result of the educational experience (transformational learning), adult learning in different contexts, such as distance, informal and self-directed learning, as well as different active or experiential instructional approaches, such as problem-based learning (Merriam, Caffarella, & Baumgartner, 2007).

For this work the adult learning concepts that have been be applied to a transferable instructional model (as opposed to teacher-specific traits like creating an environment of respect) are mutual self-diagnosis, collaboration, and meaningful work that is directly and immediately applicable to the student. The elements of collaboration and meaningful work are similar to those described in the previous section on constructivism. Mutual self-diagnosis overlaps with concepts in meta-cognition, which is addressed in more depth in the reading scientific text section.

**Science literacy theory and information literacy theory.** Constructivism and adult learning theory are the theoretical frameworks shaping the instructional strategies
for the model. Science and information literacy theories form the theoretical framework for the skills and knowledge to be delivered through the instructional model.

Being able to read, evaluate, interpret and clinically apply medical literature to medical practice is a specialized form of both science literacy and information literacy. Science literacy has a variety of definitions, from general and public oriented to detailed and context oriented. In the book *Science Matters: Achieving Scientific Literacy*, Hazen and Trefil (1991) presented a public-oriented definition, which is the scientific knowledge people need to know to understand public issues. Showalter (as cited in Laugksch, 2000) presented a more thorough definition of science literacy stating that a person who is scientifically literate is someone who:

- understands the nature of scientific knowledge;
- accurately applies appropriate science concepts, principles, laws, and theories in interacting with his universe;
- uses processes of science in solving problems, making decisions, and furthering his own understanding of the universe;
- interacts with the various aspects of his universe in a way that is consistent with the values that underlie science;
- understands and appreciates the joint enterprises of science and technology and the interrelationship of these with each and with other aspects of society;
- developed a richer, more satisfying, more exciting view of the universe as a result of his science education and continues to extend this education throughout his life;
• has developed numerous manipulative skills associated with science and technology.

Most of these general descriptors apply to the scientifically literate medical practitioner. As the thinking around science literacy grew, different types of science literacy were identified, including methodological, professional, universal, technological, journalistic, and science policy literacy (Branscomb, as cited in Laugksch, 2000). Even within these categories are subcategories, like chemistry, biology, and medicine.

In addition to its connection with science literature, EBM also relates closely to information literacy. The American Library Association (ALA) has developed competencies in information literacy which include:

• knowing the variety of types and formats of potential sources;
• accessing information effectively and efficiently;
• reevaluating the nature and extent of the information need(ed);
• accessing needed information effectively and efficiently;
• evaluating information and its sources critically;
• incorporating selected information into his or her knowledge base and value system; and
• using information effectively to accomplish a specific purpose (retrieved from http://www.ala.org/ala/mgrps/divs/acrl/standards/informationliteracycompetency.cfm).

Each of these ALA competencies aligns closely with the EBM components introduced in Chapter 1: (1) accurately assessing a patient; (2) asking appropriate clinical questions to implement a meaningful literature search; (3) acquiring a body of high
quality literature; (4) appraising the literature on validity, importance and usefulness (which is the focus of this dissertation), and; (5) interpreting and applying the evidence to the clinical issue (also the focus of this dissertation) (retrieved from JAMAevidence.com). As mentioned previously, each component has its own set of skills and knowledge to achieve competency. But for this dissertation the focus is on appraising, interpreting and applying the literature.

**Instructional Approaches: Scaffolding, Science Instruction and Reading Scientific Text**

There are numerous instructional approaches to consider in developing an instructional model for science-oriented learning. Instructional approaches for the experimental elements of science will differ from approaches teaching the complex, knowledge-intensive side of science. Scaffolding has repeatedly been demonstrated to be more effective in science instruction where there is complex, layered, and vocabulary intensive learning and that is the focus of this model. The types of scaffolds and their frequency of use within instruction also appear to impact the extent of learning. For example, Brunvand and Fishman (2007) found that explicit instructional scaffolding strategically placed throughout the instruction is more effective in enhancing student learning compared to minimal scaffolding. Chen and Bradshaw (2007) discovered that in knowledge-intensive learning situations knowledge-integration prompts provided better support for student learning compared with more open-ended problem-solving prompts.

Additional scaffolding techniques that have been found to be effective use cognitive and metacognitive guided questioning or prompts to help integrate knowledge by comparing ideas, identifying connections, distinguishing characteristics between
concepts, and identifying valid relationships (Davis & Linn, 2000; Chen & Bradshaw, 2007). Blank (2000) showed that oral discourse on the plausibility and utility of scientific results and knowledge claims resulted in significantly better retention compared to the control group that used “normal” seventh grade learning cycles.

In their research, Davis and Linn (2000) used different types of prompts (self-monitoring versus activity prompts) that encouraged knowledge integration and autonomy by “modeling reflection and providing an explicit place for reflection.” Self-monitoring prompts “encouraged students to reflect on their own learning” (e.g. elements they did not understand or encouraging thinking around how concepts fit together) and activity prompts had students reflect on activity steps, process and progress (e.g. “The major claims made by the article include…”) (p. 824). Davis and Linn found that the activity prompts “were more successful at directly eliciting scientific ideas” (p. 830) and the self-monitoring prompts “elicited better explanations” (p. 831) and connections between ideas. They concluded that the presence of both types of prompts was important, but so too was the balance of the prompts. For example, too many activity prompts reinforced “efficient step-by-step responses rather than the integrated understanding we desire” (p. 834), and more frequent self-monitoring prompts elicited an integrated understanding.

Scaffolding techniques also appeared to be helpful in teaching students reading comprehension from science texts. Mastropieri, Scruggs, and Graetz (2003) identified several scaffolding techniques for successful reading comprehension instruction:

- utilize a combination of cognitive strategies, direct instruction, and guided and independent practice;
• ask and answer questions about text while reading; and
• utilize summarization strategies.

Scaffolds also need to consider text-specific elements to help improve reading comprehension. Mayer (1975) and other researchers demonstrated that questions or prompts in earlier passages of the text impact the type of attention a reader gives and impacts reading comprehension (Mayer, 1975; McConkie, Rayner, & Wilson, 1973; McGaw & Grotelueschen, 1972). This concept can extend to teachers presenting questions for students to consider while reading (Sagerman & Mayer, 1987). Mastropieri, Scruggs, Bakken, & Whedon (1996) showed that challenging students with analytical questions and meta-cognitive activities during reading that drew them deeper into the text, almost doubled reading comprehension.

As found with other instructional scaffolding research, the location and type of questions or prompting within a text can impact reading comprehension. For example, when questions and prompts are imbedded early in the text, comprehension improves (Mayer, 1975; McConkie, Rayner, & Wilson, 1973; McGaw & Grotelueschen, 1972). Sagerman and Mayer (1987) showed that the types of questions asked during reading were important. In their study students read a series of four passages and, after the first three passages, they were presented with either verbatim questions, conceptual questions, or no questions (control). The students then read the fourth passage and answered both verbatim and conceptual questions about the final passage. Students who were prompted with verbatim questions showed improved performance on verbatim tests, but did not perform as well on conceptual tests. Interestingly, students who were prompted with conceptual questions showed improved performance on both types of tests but excelled
on the conceptual tests. Both the verbatim and the conceptual groups did better than the control, with the conceptual group scoring more than 70% higher than the control group on conceptual questions, and the verbatim group scoring 40% higher than the control group on verbatim questions. When students were prompted with content-related conceptual questions during reading (i.e. questions that ask students to extend and apply knowledge they have just learned into a new situation) there was statistically significantly improved performance on both verbatim and conceptual tests.

Although there are specific scaffolds described in the science reading comprehension literature, such as question prompts, there are other important contributions to reading comprehension. For example, a reader’s background knowledge, the text structure, text organization and coherence, and vocabulary are all important factors in reading comprehension (Kintsch, 1988, 1992). The reader’s awareness of and familiarity with how a text is organized, including how and where main ideas are stated, and subsections within texts, improves reading comprehension (Cook & Mayer, 1988; Taylor & Beach, 1984). This appears to especially be true with unfamiliar topics (Taylor & Beach, 1984).

Text coherence can refer to several aspects of the text, including the flow of text, the use of meaningful connecting words, such as “therefore” (which implies cause and effect), and the inclusion (or exclusion) of background information. McNamara, Kintsch, Songer, and Kintsch (1996) showed that novice readers, even at the college level, required coherent text with in-depth explanations and clear transitions to support reading comprehension and that text coherence in the form of supplying background information, even in the form of one or two sentences, improved reading comprehension. The
additional background information provided context and is situational, and, according to Kintsch, reading comprehension is maximized when a good situational model is linked with the reader’s long-term knowledge (Kintsch, 1994).

Vocabulary level and frequency of new terms and concepts in the text have been demonstrated to impact reading comprehension. Although the meaning of new words in more general reading situations can often be derived from the surrounding text (Mezynski, 1983), this may not be true when reading advanced science texts such as scientific/medical literature, which are vocabulary intensive and conceptually abstract. In vocabulary-dependent situations, research shows that a student’s vocabulary size and depth can impact reading performance. The larger the vocabulary repertoire the better the reading performance (Quian, 2002).

A challenge with teaching students how to evaluate medical literature is navigating the inherent challenges in simply reading medical literature. As mentioned previously, text coherence along with extensive new vocabulary in primary scientific literature can be problematic. This is coupled with the layered challenges of determining research quality. Thus, in addition to considering the research on teaching science and reading in science, it is important to look at specific research and publications that have explicitly incorporated primary scientific literature into instruction.

Publications Describing the Incorporation of Primary Scientific Literature in Instruction

This section of the literature review presents a summary of the publications that describe the incorporation of primary scientific literature (i.e. scientific research published in a professional journal – medical literature is a type of primary scientific
literature) in instruction. Most of these instruction-related publications are from work at the college level from a variety of sciences (e.g. biology, pharmacy, geology), but some occur in high schools. Reviewing these publications provided some insight into the instructional approaches used and data associated with the related research.

Many of these works provided basic descriptions of lessons and instructional approaches. Except for using the broad description of “active learning” as an educational premise, the work was typically not presented in the context of educational theory or more specific, evidence-based best instructional practices that consider student learning. Blank (2000) deemed the over-use of active learning as “activitymania” and posited that too often teachers and students equate learning with doing and that “any lack of understanding just requires another activity” (Blank, as cited in Yore & Treagust, 2006). Many of these publications also did not present direct evidence of student learning beyond anecdotal observations or student self-assessment or opinion surveys about their perceived improvements in confidence and ability.

There are several key publications (Yarden, Brill, & Falk, 2001; Brill, Falk, & Yarden, 2004; and Baram-Tasabari & Yarden, 2005) that presented substantive and interesting data directly related to teaching and student learning when using primary scientific literature in instruction. These publications informed a significant portion of the approaches used in the instructional model for this work.

**Typical publications in teaching using primary literature/EBM.** As mentioned previously the majority of publications on using primary scientific literature in instruction were descriptive and not done in a rigorous research context. However, these publications are still useful to consider and can provide ideas and indications about how to approach
instruction. Karcher (2000) provided a very brief description of a primary scientific literature review activity she did with a sophomore-level genetics and molecular biology class of 200 students. She did not frame her work within basic learning or instruction constructs nor did she collect data. She did list some “problems” she encountered and possible solutions she would try in the future. Although not explicitly stated, the activity is assumed to be individual rather than group. Karcher admits that she did not “give the students many guidelines” (p. 485). The activity was conducted as follows: students first emailed their list of primary scientific literature they wanted to summarize and got pre-approval; they wrote 1-2 paragraph summaries of the articles; the instructor evaluated the summaries and students had an opportunity to re-write them if needed; the summaries were posted on the course website. This appears to be a poorly scaffolded activity and the problems she encountered reveal this weakness. Many of the solutions Karcher proposed are examples of adding scaffolding to the instruction, such as providing more examples, giving the students “more guidelines on ‘how to write’ a summary paragraph,” and asking the students “to include a second paragraph explaining the relation between the article chosen to summarize and the course material” (p. 486).

DebBurman (2002) published a more detailed and informative description of five instructional activities she conducted in her sophomore level introductory cell biology courses to “promote process-skill development within content-rich pedagogy and to connect text-based and laboratory based learning with the world of contemporary research” (p. 154, 155). The educational learning and instructional constructs upon which she based this work included “emphasizing process skill development versus content-based teaching” (p.154), using active learning, cooperative learning, collaborative
learning, and experiential learning. Although she did not explicitly mention scaffolding as an instructional approach, the activities were highly scaffolded with detailed support elements, such as specific lessons and materials on how to conduct a literature search. In addition, concepts and skills built upon one another in the activity.

DebBurman utilized a mix of small group work and large group interactions, individual work and lab experiments, with written summaries and presentations as products. All of this was done within the context of the class as a scientific community that mimicked the communication and peer structures in real-world science. The series of activities included:

1. Students worked in groups and became familiar with one primary scientific article from a leading peer-reviewed journal.
2. They presented the article and discussed them via Power Point based journal clubs.
3. Students individually wrote a journalism-type of summary of a medical discovery from a single primary article.
4. Then students worked in groups to investigate the molecular basis of a disease using several primary scientific articles and presented their work in a “disease symposium seminar” that included the class and invitees.
5. Students wrote a lab report in the form of primary scientific literature which became part of an “in-house journal.”

A pre- and post-survey was conducted to measure student attitudes towards learning (n=98). DebBurman summarized the following conclusions from the survey results: Students perceived that they improved in their ability to communicate
contemporary research; Students rated four of the five activities highly stating that they were relevant and met course goals, the news journalism activity received the lowest ratings. An interesting survey result that was not expanded upon was that students did not perceive improvements in reading primary scientific literature.

McNeal and Murrain (1994) incorporated primary scientific and medical literature in their undergraduate drug information class. They situated their work in cooperative learning and what they described as “constructivist approaches” which is in the form of student-initiated inquiry using project-based approaches where students “must ask their own questions.” However, when reading through the details of the activity it appears some instructional scaffolds were provided, but not detailed (the authors did not present their work in terms of scaffolding). Scaffolds included: having students write a summary of the secondary literature to research a drug of choice (which develops library search skills, topic development, and acquaints students with vocabulary); building the “skills necessary to critically approach primary scientific articles” by providing a pre-selected article and guiding questions to reading and evaluating the article; and having a group report on a primary research article. All that was completed prior to doing the individual project synthesizing primary literature on a topic of their choice.

The authors did not present any data, but made statements like “students learn how to ask testable questions” (p. 4 of online version). The authors did not describe how they reached such a conclusion, nor did they describe whether explicit instruction, guidance, or feedback was provided for students to develop this skill.

There are two publications by a team highlighting work from the same course, a Doctor of Pharmacy second-year Drug Information and Literature Evaluation course.
The first publication, Timpe, Motl, & Eichner (2006), presented a brief summary of the weekly learning activities and survey data on student self-assessing changes in their abilities and confidence. The authors briefly situated the work in the educational context of active learning, but not beyond a short statement. A list of nine learning activities was provided with few implementation or instructional details. Two of the learning activities utilized medical literature which were, “apply critical literature evaluation skills using two original research articles” and “present a journal club” (p. 2). The first medical article given to the students was used by the instructor to explain the concept of literature evaluation (this is a scaffold but it is unknown if additional scaffolds in the form of questions or rubrics were provided) again the authors did not present their work in terms of scaffolding.

The students practiced the literature evaluation with the second medical literature article. Students worked individually but engaged in class discussions (a potential scaffold depending on the nature of the discussions), then presented in small group sessions for journal club.

Pre- and post-survey results showed increases in students’ self-rated ability and confidence in all areas. Ninety-four percent of the students perceived they could accurately describe the study design and 96% perceived they could assess the external validity of the study (n=69). This is compared to 58% and 61% respectively, self-reporting their level of confidence in a pre-survey (n=76). The students reported comparable improvements in their ability to critically evaluate methodology (50% pre-survey to 86% post survey), analyze the statistical analysis of the data (54% to 75%) and identify ways to improve the research design. Only 61% of the students agreed or
strongly agreed that the journal club experience was valuable and met learning objectives. This may be a function of how the journal club was implemented (a discussion of journal clubs is provided in the next sub-section).

No statistical analyses were performed to determine significance of the changes, and no additional assessments were reported to determine whether students' self-perceptions accurately reflected ability.

Motl, Timpe, and Eichner (2006) also described an activity where students individually read one of six pre-selected medical journal articles and a health news publication based upon that article and evaluated the accuracy of the health news publication. Motl, et al. very briefly situated their work in the context of developing the EBM skills of literature retrieval and evaluation and the educational construct of active learning. They conducted a survey (n=98) of student perceptions of skills developed and the activity’s contribution to achieving the course objectives.

The activity was delivered as follows: students were assigned one of six health-related news articles; they conducted a literature search to retrieve the medical literature upon which the article was based; students read and evaluated the medical literature “using the knowledge and skills gained throughout” the course (p. 2); students then evaluated the “quality and accuracy of the news publications’ representation of the original research” (p. 2); students wrote a three- to five-page written critique of the news publication. Although the authors did not frame their work in terms of scaffolds, several scaffolds were provided to support student learning. These included guiding questions (e.g. Did the news article provide sufficient and accurate background information
compared to the original research article?) and students were provided a copy of the grading rubric the instructors used to evaluate the students’ critiques.

Pre- and post-survey data (n=98) provided self-rated ability and confidence, which improved for all measures (increases of 4%-32%; the measures were ability and confidence in the areas of searching for original research, evaluating accuracy of news publications, identifying limitations of news articles, and identifying potential questions from the lay public). No statistical analyses were performed.

**Journal clubs.** The majority of the publications focused on teaching students about primary scientific literature were related to the use of journal clubs. Again, data presented with most of the articles were satisfaction surveys and students self-perceptions of changes in ability and/or confidence. There was very little, if any, situating the work in terms of instructional best practices or student learning paradigms, such as adult learning, or using direct measures of learning.

Burstein, Hollander, and Barlas (1996) developed a “structured review instrument” (SRI) for a journal club for emergency medicine residents. The club met once a month as part of the didactic curriculum for residents (on average 10-12 of the 18 residents attended the sessions) reading approximately 75% of the assigned articles. Five medical articles were reviewed, three of which were pre-selected and the remaining two were selected based upon participant interests. The residents signed an attendance sheet and indicated if they read the article ahead of time (which was required) and then one of those residents who read the article was selected to begin the review/discussion. Following the discussion the instructor and/or an attending statistician summarized “the lessons to be learned both in terms of study design and clinical practice” (p. 3).
The SRI was a checklist that provided more structure to the article discussion prompting the journal club participants to discuss the hypothesis, methods, study population, measurements, outcomes, variables, flow, statistics, results, conclusions and limitations. Previously the approach to the article discussion was dependent on the resident leading the discussion, which varied with each person.

The authors did not present their work within an educational teaching or learning construct, nor did they describe the development of the SRI. Instead they very briefly described the prevalence and importance of journal clubs in medical education. The authors administered a brief satisfaction survey that showed increased satisfaction with the journal club after the introduction of the SRI (which is a scaffold).

Lee et al. (2006) implemented a “systematic checklist of review criteria for analyzing journal articles” then conducted a pre- and post-survey where residents self assessed their ability in five domains: “appraise and assimilate evidence; read a journal article critically; use a systematic and standardized checklist; apply knowledge of study designs and statistical methods; and maintain a self-documented written record of compliance” (p. 498). The majority of the article discussed the survey results, which showed statistically significant improvements in self-assessment scores following the journal club experience. The authors framed their work via accreditation requirements for medical education in the areas of practice-based learning and evidence-based medicine. They did not present the work within formal teaching and learning constructs, nor did they describe the development of their checklist.

There are many other short articles in medical education journals that provide general, very simple descriptions of approaches to journal club including: Alguire, 1998;

The pharmacy and medical education literature presented various and comparatively informal instructional tools and approaches for how to train students to read and critically evaluate the medical literature (Kollar, Fischer, & Slotta, 2005; Linn, Clark, & Slotta, 2003; Collins, Brown, & Newman, 1989; Brunvand & Fishman, 2007). For example, the most commonly described instructional support tool in medical education publications is the use of stand-alone questions to teach students to evaluate medical literature (Kollar, Fischer, & Slotta, 2005; Brunvand & Fishman, 2007). Although the use of stand-alone questions can provide some structure for students when learning to evaluate medical literature, weaknesses exist in that approach.

In order for students to answer such questions they must either possess extensive background knowledge or depend highly upon the faculty or an expert to provide explanations. In addition, stand-alone questions in and of themselves do not necessarily teach students how to judge the quality of a study nor its associated publication, leaving students again to rely heavily on faculty. Having faculty/experts present to support student learning is important and necessary, and arguably irreplaceable, however, utilization of faculty expertise should be strategic and combined with explicit and targeted instructional support to build solid, transferable skills in students, particularly in complex learning situations.

Publications in teaching using primary literature/EBM that consider education research and theory. A group of researchers from Israel have contributed
substantively to informing instructional methods to teach students how to read scientific literature. Yarden, Brill, and Falk (2001) and Brill, Falk, and Yarden (2004) approached their investigations into cognitive and instructional-related research in reading scientific literature with high school students. Through two different studies, Yarden et al. (2001) and Brill et al. (2004) connected cognitive and instructional methodological processes to advance student understanding of the information contained in published scientific studies. Much of their work used and provided additional evidence for the use of scaffolds, like strategically-placed questioning, discussion, and making the organizational structure of the text explicit to advance scientific literature reading comprehension.

Yarden et al. (2001) developed a “conversational model for learning through research articles” in which high school students read scientific literature “together in the classroom one section at a time” (p. 192). At the end of each section of the study, students were asked clarifying or research questions and the teacher wrote down the questions for all students to see. Students then attempted to answer some of the questions by offering hypotheses or making outcome predictions about the experiment presented in the scientific paper. As the class continued to read through the sections of the study, they revisited their questions and revised their answers. During this process the teacher moved from a leading role (i.e. modeling the reading, questioning, and answering process) to a moderator, encouraging the generation of questions from the students.

The authors evaluated the process students used to understand the scientific literature by (1) monitoring the types of questions asked by students, (2) observing which students were asking the questions (low- versus high-ability students), and (3) analyzing the scientific conversations that occurred between students and the class. They found that
most of the questions asked by students, particularly at the beginning of the reading, were information-gathering questions (e.g., definitions) and students of all skill levels were actively engaged in the process. Once students grasped the content knowledge, they engaged in debates and spontaneously demonstrated advanced, abstract thinking by suggesting things like a new experimental design to answer some of the questions raised during the reading process. This observation appears to support the concept about the importance of establishing a knowledge base, including vocabulary, before being able to expand into analysis.

Brill et al. (2004) learned more about how students read scientific literature and how scaffolds can support reading, by observing two intermediate-level high school students reading a “minimally-modified” developmental biology research article. Initially the students read through the article and tried to interpret it on their own. They were then given questions related to the article to further guide the reading. The authors observed the pair of students as they conferred with each other and processed the information. The observers were able to note revelations about the students’ discoveries finding that the students initially encountered comprehension difficulties. This was attributed to the “lack of schemas and automation that expert readers possess” (p. 511). The authors referred to the students as having a “fragmented cognitive structure” where a lack of experience (and the knowledge obtained by that experience) resulted in insufficient schemas through which to process the information.

However, once the students grasped the “classical structure” of the research article (text organization) and were given questions to help them focus on specific aspects of the article (prompts), the students made compelling cognitive leaps. For example, for
most of the session the students misunderstood the conclusions from the research. However, once the students began re-reading the text to answer the questions, their understanding of the article deepened and they independently realized their initial interpretations of the conclusion were incorrect.

Baram-Tasabari and Yarden (2005) expanded on this work in text structure in scientific literature and found that students who read secondary literature (e.g. popular science articles or textbooks), with its coherent text and flow, showed improved comprehension of the content compared with students who read primary scientific literature covering the same content, which is much more fragmented with poor flow (fewer transitions and less background information). Interestingly, though, the students who read the primary scientific literature, demonstrated better inquiry skills and had a comparatively expanded understanding of the scientific process.

Challenges the authors faced included: reading and evaluating original research can be time consuming because of the students’ inexperience at reading medical literature; choosing the medical literature because of limitations in the student’s knowledge base; finding appropriate medical literature that corresponds with news articles; and the activity assessments were time consuming to grade.

Blommel and Abate helped the EBM education field take a step forward by being one of the first to develop (and publish) a scoring rubric for journal club presentations in a Doctor of Pharmacy curriculum. This is important because prior to Blommel and Abate many medical literature evaluation instruments were simple checklists that provided very little information and tended to be yes/no or present/absent, rather than considering important grey areas in between.
The Blommel and Abate rubric incorporated elements of the CONSORT statement for reporting of randomized control trials (Moher et al., 2001) (CONSORT is discussed in more detail in the next section on EBM). An abbreviated checklist was provided for students showing them what to present during journal clubs and elements of what to evaluate when reading the medical literature. Students checked for the presence of specific information in the medical literature (e.g. the presence of blinding in a study) and then assigned a score from 0-2 in one of three major areas in the rubric (the study design, outcomes and data handling, and statistics and conclusions). A score of 0, meant the authors did not accurately and completely report the information for the area, 1 point meant most of the information reported was accurate and complete, and 2 points meant all of it was accurate and complete.

Although the Blommel and Abate rubric was an improvement over open-ended questions and advances instruction in this area, it had several weaknesses. First, it lumped elements of the article together into just three broad areas, essentially ignoring the organizational structure of the medical literature text. As described earlier in this paper, the educational research on improving science reading shows knowledge of text structure is important in reading comprehension (Cook & Mayer, 1998; Taylor & Beach, 1984), thus a rubric should more closely follow to text structure to help support student comprehension.

A second weakness of the Blommel and Abate rubric is that it was only a checklist for inclusion of information in the publication, it did not define or teach students important literature evaluation terms. Although the rubric scale mentioned accuracy, no information was provided to students to help them determine accuracy. For
example, the Blommel and Abate rubric had a single check box for “blinding” placed under their broad categorical area of design (i.e. asking “is blinding present or not?”), but there are multiple levels of blinding (single, double, and triple) that can occur and impact the quality of a clinical study. Results from the pilot research for this dissertation further supports these conclusions (Dawn et al., 2011), thus, if poor quality or inaccurate clinical conclusions are reached, these skills could be carried over as practitioners.

Evidence Based Medicine Literature

Literature to inform the development of the Medical Literature Evaluation (MLE) Rubric. In addition to using the educational literature to develop the instructional model, the EBM skills and knowledge developed by the instruction must also be informed by the EBM literature. Over the years EBM practitioners and advocates have been identifying and defining the factors that constitute a high quality study and publication of that study. Some of the earliest publications on the topic defined a few simple parameters to consider, such as asking if the study design is appropriate to the objectives, identifying elements of a representative sample, and identifying components of an acceptable control group (Fowkes & Fulton, 1991). More recent publications have increased the level of sophistication in identifying factors that can increase bias or lead to erroneous conclusions.

For example, the Consolidated Standards for Reporting Trials (CONSORT) (Moher et al., 2001, 2010) was (and still is) a seminal work in this area. It was developed for biomedical journals and authors to clearly define study elements that need to be included the publication of clinical trials. The CONSORT was the primary work used to guide the development of the MLE Rubric, an important scaffold in the instructional
model for the pilot study and this dissertation. CONSORT addresses the following topics via a checklist (Table 1) and provides more detailed descriptions of each topic in a 37-page paper (note that spellings in the table are those used in CONSORT).
Table 1

CONSORT 2010 Checklist of Information to Include when Reporting a Randomized Trial

<table>
<thead>
<tr>
<th>Title and abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a Identification as a randomised [British spelling] trial in the title</td>
</tr>
<tr>
<td>1b Structured summary of trial design, methods, results, and conclusions</td>
</tr>
</tbody>
</table>

**Introduction**

- Background and objectives
  - 2a Scientific background and explanation of rationale
  - 2b Specific objectives or hypotheses

**Methods**

- Trial design
  - 3a Description of trial design (such as parallel, factorial) including allocation ratio
  - 3b Important changes to methods after trial commencement (such as eligibility criteria), with reasons

- Participants
  - 4a Eligibility criteria for participants
  - 4b Settings and locations where the data were collected

- Interventions
  - 5 The interventions for each group with sufficient details to allow replication, including how and when they were actually administered

- Outcomes
  - 6a Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
  - 6b Any changes to trial outcomes after the trial commenced, with reasons

- Sample size
  - 7a How sample size was determined
  - 7b When applicable, explanation of any interim analyses and stopping guidelines

**Randomisation**

- Sequence generation
  - 8a Method used to generate the random allocation sequence
  - 8b Type of randomisation; details of any restriction (such as blocking and block size)

**Allocation concealment mechanism**

- 9 Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned
<table>
<thead>
<tr>
<th><strong>Table 1, continued</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Implementation</strong></td>
</tr>
<tr>
<td>10 Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions</td>
</tr>
<tr>
<td><strong>Blinding</strong></td>
</tr>
<tr>
<td>11a If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how</td>
</tr>
<tr>
<td>11b If relevant, description of the similarity of interventions</td>
</tr>
<tr>
<td><strong>Statistical methods</strong></td>
</tr>
<tr>
<td>12a Statistical methods used to compare groups for primary and secondary outcomes</td>
</tr>
<tr>
<td>12b Methods for additional analyses, such as subgroup analyses and adjusted analyses</td>
</tr>
<tr>
<td><strong>Results</strong></td>
</tr>
<tr>
<td>Participant flow (a diagram is strongly recommended)</td>
</tr>
<tr>
<td>13a For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome</td>
</tr>
<tr>
<td>13b For each group, losses and exclusions after randomisation, together with reasons</td>
</tr>
<tr>
<td><strong>Recruitment</strong></td>
</tr>
<tr>
<td>14a Dates defining the periods of recruitment and follow-up</td>
</tr>
<tr>
<td>14b Why the trial ended or was stopped</td>
</tr>
<tr>
<td><strong>Baseline data</strong></td>
</tr>
<tr>
<td>15 A table showing baseline demographic and clinical characteristics for each group</td>
</tr>
<tr>
<td><strong>Numbers analysed</strong></td>
</tr>
<tr>
<td>16 For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups</td>
</tr>
<tr>
<td><strong>Outcomes and estimation</strong></td>
</tr>
<tr>
<td>17a For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)</td>
</tr>
<tr>
<td>17b For binary outcomes, presentation of both absolute and relative effect sizes is recommended</td>
</tr>
<tr>
<td><strong>Ancillary analyses</strong></td>
</tr>
<tr>
<td>18 Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory</td>
</tr>
<tr>
<td><strong>Harms</strong></td>
</tr>
<tr>
<td>19 All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)</td>
</tr>
<tr>
<td><strong>Discussion</strong></td>
</tr>
<tr>
<td><strong>Limitations</strong></td>
</tr>
<tr>
<td>20 Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses</td>
</tr>
</tbody>
</table>
Although very useful in its content to guide journal editors and authors, the CONSORT Statement is not best suited for use with students. This is because of the background knowledge required to understand the concepts and terminology in the document and the document formatting is either a too short checklist or a too long 37-page paper. Students need a simple, well-formatted tool, like a rubric, that can serve as a guide providing basic definitions and parameters for evaluating medical literature as they are reading.

Like the medical education community, the professional biomedical community has developed and published a number of clinical literature evaluation tools, which are also primarily in the form of checklists. These checklists are used by practitioners and developers of clinical guideline or systematic reviews, educators sometimes use these instruments in their instruction as well. The challenge with these checklists is that most of the people using them already have a grasp of the vocabulary and concepts and are experienced with reading clinical literature. Many of these checklists are highly simplistic with yes-no indicators for the presence or absence of certain information or with scales like “criterion entirely fulfilled, criterion mostly fulfilled, criterion mostly not fulfilled, criterion not at all fulfilled, criterion not described adequately, criterion not applicable” (Liddle, Williamson, & Irwig, 1996) without any in-depth descriptors about what constitutes terms like “mostly fulfilled.” Students who do not have the experience with conducting research or reading research studies have nothing to base such decisions on without explicit guidance.

Most of the existing checklists are designed for a specific type of research called a randomized control trial (RCT), also called a clinical trial. The reality of practicing
evidence-based medicine is that other types of studies, such as case-control and cohort, will likely be included in literature reviews when students investigate the state of the science on a particular topic. Although there are some similarities between the elements that constitute a high quality clinical trial and a high quality case-control or cohort study, there are also differences. Thus, evaluation rubrics are needed to explicitly guide students on specific features of those types of studies.

The CONSORT was used to guide the development of the initial versions on the CLE Rubric used in the pilot research and the current version of the expanded MLE Rubric used for this research. Additional resources that were used to guide the development of the MLE Rubric (which considers RCT, case control, and cohort studies) are provided below. Again, many of the existing documents, rubrics or checklists are useful as guides but still lack descriptive detail and evaluative guidance needed for students and likely could not be used “as is” for the activity associated with this dissertation.

- Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies (von Elm et al., 2007).
  - Similar to the CONSORT, the STROBE statement provides recommendations for journal editors and authors regarding what should be included in an observational study (which includes cohort, case-control and cross-sectional studies).
Developed during a 2-day workshop in Bristol, UK in 2004 with 23 attendees representing journal editors and scientists from Europe and North America.

The authors stated that “the STROBE statement was not developed as a tool for assessing the quality of published observational research” (p. 1456).

- **Transparent Reporting of Evaluations with Non-randomized Designs (TREND) statement** (Des Jarlais, Lyles, Crepaz, & the TREND Group, 2007).
  - Similar to the COSORT and the STROBE statements, TREND provides recommendations for journal editors and authors regarding “standardized and transparent reporting” of non-randomized intervention studies in behavioral and public health.
  - Developed at the Centers for Disease Control Journal Editors Meeting in July 2003.

  - Developed by the Centre for Clinical Policy and Practice of the New South Wales Health Department, Australia.
  - Provides a standardized approach to assist professionals who are reviewing literature to develop clinical or health-related guidelines.
  - Provides five different checklists to allow reviewers to assess study quality (reviews of the effect of interventions, studies assessing the effect of interventions, interrupted time series studies assessing the
effect of interventions, studies assessing risk factors, studies assessing diagnostic accuracy).

- Uses a coding method for evaluating criteria (criterion entirely fulfilled, criterion mostly fulfilled, criterion mostly not fulfilled, criterion not at all fulfilled, criterion not described adequately, criterion not applicable).

  - Presents questions for the reviewers to help determine whether the guidelines or recommendations they made are “valid and likely to benefit a population” (p.1).


  - A series of simple checklists with questions and brief statements intended to guide the evaluation of the quality of systematic reviews and meta-analyses (Checklist 1), controlled trials (Checklist 2), and cohort studies (Checklist 3).

**Synthesizing research to reach a clinical conclusion.** After students have learned what constitutes a high quality study by using the medical literature evaluation rubrics, they need to learn how to reconcile information from multiple studies, multiple types of studies, and studies that may reach differing conclusions.

The literature that informed the development of this feature of the instruction are from guidelines for systematic reviews and academic discussions about research synthesis and comparability work. Themes that emerged in these conversations about
research synthesis and comparability are highlighted below and were considered in the design of the lesson:

- methodological differences between studies and how those can be reconciled when comparing studies and their results;
- topic diversity between studies, which refers to the common occurrence that even if two studies investigated a similar topic, exact overlap of the research question is unlikely;
- diversity in the study details and whether to treat them as one variable or as individual variables, and the associated advantages and disadvantages with each;
- terminology issues including using different words to denote similar concepts (e.g. medication adherence versus medical compliance versus patterns of use) and conversely using common terminology differently [e.g. referring to a study design as a randomized control trial when it is actually more like an observational study (Sandelowski, Voils, & Barroso, 2006)];

The ability to compare and weigh the quality of two or more studies on the same research topics and draw clinical conclusions from multiple and different kinds of studies requires in-depth knowledge of the medical topic as well as research methodologies within a given research type. Methodologies include things like sampling methods, sample numbers requirements, inclusion/exclusion criteria, blinding, and other ways systematic bias and/or error can be addressed or minimized through the study design. It also includes how data can (and should) be analyzed using appropriate statistical methods and reasonable interpretation of statistics for that study type. In addition to the need to
understand the research methods, medical literature evaluation requires knowledge of the medical research topic (e.g. the mechanism of action of a specific drug on a specific disease state) and the vocabulary and procedures (such as common lab tests) associated with that medical issue.

The purpose of detailed medical literature evaluation rubrics that are focused on specific research designs (RCT, case-control and cohort) is to teach the students the specific elements that constitute high quality studies and allow students to interact with and develop informed judgments about a study they read. In the context of considering multiple studies on the same topic, the medical literature evaluation rubrics can help students identify studies that should probably be excluded from consideration or provide a basis for weighting studies in reaching clinical conclusions. For example, one consideration is whether the results of a high quality cohort study “outweigh” a poor quality clinical study even though the clinical study design is the “gold standard” and is higher on the hierarchy of evidence (Figure 4). To do this, students need to learn to step back from the detailed view provided by the evaluation rubrics and operate from a bird’s eye view while still focusing on appropriate key elements.
The hierarchy of evidence provides medical practitioners with general guidelines about the quality of evidence of different types of medical research. This is intended to help clinicians make medical decisions based upon the highest quality of evidence. (retrieved, http://www.mclibrary.duke.edu/subject/ebm/ebmpyramid.html)

When reconciling methodological differences in studies it is important to identify the least common denominators of the studies. First, when trying to evaluate the quality of the study a least common denominator is validity, which means that the study methods “can be trusted to provide a genuine, accurate account of the phenomenon being studied” (Avis, 1994). Validity can be threatened by flaws in subject selection, poorly defining and poorly recording information, and either flaws in the study design or execution (Streiner et al., 1989 as cited in Avis, 1994).

For example, three primary factors are measured in the RCT, cohort and case control study designs and can be considered in comparisons - the intervention, potential confounders, and the outcome (Downs & Black, 1998). Another commonality is that each
study type is vulnerable to some of the same biases. Bias can be defined as “anything besides the treatment or random chance that can modify the strength or direction of the association between treatment and outcome” (Doig, 1998, p.514) and specific types of systematic error/internal validity (bias) include selection bias, performance bias, detection bias, attrition bias (Downs & Black, 1998; Juni, Altman & Egger, 2001).

Methods to reduce bias can vary with study design and can include blinding and “selection of unambiguous and appropriate end points,” as well as random assignment to the intervention and “complete follow-up and reporting on all patients recruited into the trial” (which do not occur in cohort and case-control studies) (Doig, 1998, p.514).

Literature that considers the review of different types of studies identifies common themes and associated questions to compare the studies with respect to these key elements related to validity and bias (Avis, 1994; Fowkes & Fulton, 1991; Downs & Black, 1998; Reisch, Tyson, & Mize, 1989). According to Fowkes and Fulton (1991, p.1140) an evaluation of a study can essentially be summed up in with three questions:

- Are the results erroneously biased in a certain direction?
- Are there any serious confounding or other distorting influences?
- Is it likely that the results occurred by chance?

In addition to the birds-eye themes, there are important subthemes that must be made explicit to the student for all three types of studies (RCT, cohort, and case control). These subthemes are:

- Are the research objectives/hypotheses/outcomes to be measured defined?
  - Were they defined prior to the study?
• Is the study design well matched to the research objectives/hypotheses/outcomes?

• Are the interventions clearly described?

• Is the study sample appropriate/representative?

• Is the sampling method appropriate?

• Is the sample size sufficient? How was it determined? This will be considered differently for qualitative research, for example assess the “biological representativeness” of the sample)

• Is the sample inclusion/exclusion criteria appropriate?

• Are refusals and non-respondents a source of bias?

• Is the selection and assignment of control groups appropriate? (where applicable)

• What details are provided for the treatment? (e.g. dosage, time of day, route, etc.)

• What are the sources of error? (consider the techniques of obtaining and recording information)

• What are the sources of bias?

• What quality control measures were taken? (e.g. calibration and accuracy of instruments, blinding)

• Were adverse events reported?

• Was incompleteness of the end-results (e.g. compliance, drop-outs and death, missing data) and “distorting influences” (e.g. extraneous
treatments in the sample, contamination, confounding factors, etc.) taken into appropriate consideration?

- Are the conclusions valid based upon the data?
- Is the study reproducible?

Downs and Black (1998) developed a single checklist (yes-no assessments) that assess study quality and quality of reporting, internal validity and power, as well as external validity (generalizibility) for randomized trials, case-control studies and cohort studies. A single checklist for all three study types is convenient but has challenges for development. Many of the questions in the checklist developed by Downs and Black are similar to those listed above but with some greater depth of detail. The authors developed and tested a pilot checklist for face and content validity and revised it to improve the psychometric properties (internal consistency and test-retest reliability) of the checklist.

After the revisions the authors found that “it is feasible to develop a checklist that can be used to assess the methodological quality not only of randomized controlled trials but also non-randomized studies” (p. 380) that alerts reviewers to a study’s methodological strengths and weaknesses. They stated that “there was little difference between its performance with non-randomised and with randomised studies” (p. 380). The authors did have difficulty with reliability of the external validity section of the checklist they developed, which would be one important factor in the clinical conclusion element of this dissertation work.

Downs and Black offered several possible explanations for the low reliability of the external validity section of the checklist including the small number of sub-scale questions (three), possible poor wording of the questions, and/or limited knowledge of
the reviewers in the area of external validity, and offered that the third reason is more likely. This reinforces the short-coming of many of the existing checklists with respect to having students use them, the limitation that students are not yet content or research experts.

The MLE Rubric developed for this study considers all three study types on a single rubric using the CONSORT categories, details and organization as a primary organizing approach. The Downs and Black checklist and other literature evaluation resources were used to add details and identify overlapping elements between the study types (Elm et al. 2007; Des Jarlaine, Lyles, Crepaz, & the TREND Group, 2007; Liddle, Williamson, & Irwig, 1996; Scottish Intercollegiate Guidelines Network Methodology Checklists, 2001-2009).
Chapter 3

Methods

The goal of this study was to test a scaffolded instructional model in a complex Evidence Based Medicine (EBM) lesson where students evaluated three different medical studies [randomized control trial (RCT), case-control, and cohort] on the same topic and used those studies to reach an evidence-based conclusion. The hypothesis was that the students’ literature evaluation skills and subsequent application of the literature to address a clinical question would more closely approximate the experts following implementation of the model.

Study Background

In Fall 2007 I developed and tested the first phase of an original scaffolded instructional model and an associated activity using the model. The scaffolding/activities included an original Clinical Literature Evaluation Rubric (CLE Rubric), the Key Vocabulary and Concepts Worksheet, and strategic individual, small, and large group work to teach students to critically evaluate clinical literature (component 4 of the JAMA EBM 5-core areas).

Each of the scaffolds/activity components were developed considering the elements highlighted in the literature review - text organization, vocabulary, meta-cognition, adult learning theory, and evidence-based medicine concepts. Figure 6 shows each of the scaffolding contained within the instructional model.
Scaffold one was the CLE Rubric (Appendix A), which students completed as they read the RCT. The rubric is organized following the typical text structure of clinical research - Introduction, Study Design, Statistics, Results/Conclusions - and highlights elements that need to be contained/described in each section of the study, as well as identifies what constitutes higher quality versus lower quality research and reporting (as identified by the evidence-based medicine community). This rubric applies the reading comprehension concepts of making the text organization explicit and defining clinical research-based vocabulary (e.g. allocation concealment, ancillary analyses, adverse events, etc.), as well as supports self-direction in student learning (adult learning theory).
The second scaffold had students self-identify their own knowledge gaps by highlighting vocabulary and concepts with which they were not familiar or think they needed to understand better to be able to evaluate the study. They then defined four to five terms independently and brought their definitions to share in small groups in class. This scaffold addressed vocabulary challenges to enhance text comprehension and used meta-cognition, through the act of self-identifying unknown terms, to enhance learning.

Scaffold three consisted of an interactive small and large group process to address knowledge gaps. In small groups, students shared the 4-5 terms they defined, which allowed them to compare notes on terms that overlapped, and to expand their definitions. The whole-class discussion was facilitated by the instructor to address questions and misconceptions and expand upon definitions. Students were also given a “Completed Article Vocabulary and Concepts Sheet” which contained definitions and additional information about many of the terms contained within the clinical literature, and again supported vocabulary and self-direction since not all of the terms would be covered in class.

In addition, as part of the class discussion, instructors were encouraged to ask students to reflect upon whether there were terms they thought they understood but realized they had knowledge gaps, and whether a more in-depth understanding impacted their article evaluation. This meta-cognitive exercise intended to raise student awareness of the importance of the vocabulary in literature evaluation and how their own knowledge gaps can impact accurate evaluations.

The last step in the pilot activity/model was for students to re-read and re-evaluate the pre-selected RCT using the CLE rubric. This gave students an opportunity to review
the article with increased background knowledge.

The activities for the pilot study were implemented in the second year pharmacy lab (herein referred to as lab). The hypothesis for that work was that the use of a targeted instructional scaffolding approach would improve student abilities to critically evaluate clinical literature compared to more traditional didactic instruction. Changes in student performance were assessed quantitatively and qualitatively by using pre- and post-CLE Rubric scores that students assigned to the publication, differences between pre- and post-clinical literature evaluation summaries written by the student, and by interviews/case studies. That research consisted of blinded and un-blinded elements and was determined to be exempt and approved by the HRRC in 2007 (HRRC07-316).

Results of the pilot research were that the article's overall CLE Rubric evaluation score given by the students decreased significantly (p<0.005) from a pre-activity score of 76.7% to a post-activity score of 61.7%. Changes in the rubric score indicated that students identified additional weaknesses in the article's study design (a lower article score indicates a weaker study /publication) and reporting of study data, and the authors' interpretation of that data. Additionally, the students' average post-activity CLE Rubric score was more closely aligned with the consensus score given to the article by the faculty experts (62.0%).

Conclusions from that study were that using scaffolding appeared to improve students’ ability to more critically evaluate a clinical study compared to the didactic coursework alone. This conclusion was reached because the pre-rubric scores and written evaluation summary served as an indicator of the students’ clinical literature evaluation knowledge and skills following completion of the clinical literature evaluation lecture.
material and over 80% of the in the concurrent didactic research and informatics lecture coursework was completed (which consisted of lecture, exams, and a group literature evaluation project).

Being able to evaluate a single medical study and study type (RCT) is not enough to develop an EBM practitioner. Students need to be able to read and evaluate multiple medical study types (cohort and case control in addition to RCT) and consider a body of work together to reach a clinical conclusion. This current work applies the pilot instructional model to an expanded, next-steps sequence in the lesson series (addressing components 4 and 5 of the JAMA EBM core components). I developed and tested a new lesson teaching students how to evaluate and extract key data from multiple types of medical literature (RCT, cohort and case control) on the same subject to reach a clinical conclusion, testing the instructional model in the context of this expanded instruction.

In order to measure students’ learning and identify potential modifications to the model and lesson, I conducted a descriptive mixed-methods (qualitative and quantitative) study using similar methods from my previous work.

**Study Objectives**

The goal of this study was to test the application of the scaffolded instructional model in a more complex EBM lesson where students evaluate and derive appropriate data and conclusions from three different types of studies on the same topic to reach a sound, evidence-based clinical conclusion. The study objectives were to:

- create, implement and edit a new Medical Literature Evaluation (MLE) Rubric to address case-control, cohort and RCT studies;
• create, implement and edit a new lesson teaching students how to evaluate three different studies and types of studies on the same topic, extract relevant data and reach an evidence–based clinical conclusion;
• observe implementation of the lesson in the third year pharmacy lab to note any differences in delivery between the sections and to note potential modifications to improve lesson delivery;
• develop and implement a characterization-survey (based upon a survey implemented in 2007 for the pilot study) measuring the perceptions of student confidence in reading, evaluating, understanding and applying results from multiple medical studies towards a clinical conclusion;
• develop and implement a post-activity survey asking students to rate the different instructional elements/scaffolds in the activity; and
• evaluate students’ clinical conclusions by comparing their conclusions with expert clinical conclusions.

The hypothesis was that the students’ literature evaluation skills and subsequent application of the literature to address a clinical question would more closely approximate the experts following implementation of the model.

Study Population

Located in a large public institution in the southwestern United States the College of Pharmacy (COP) has a 4-year Doctor of Pharmacy program with 84 to 90 students per class. The study population was 89 third year Doctor of Pharmacy students. The student body consisted of 55% women and 45% men and had diverse ethnicity with the following
distribution: white 48.3%; Hispanic 31.5%; Asian 12.3%; Native American 5.6%, black 2.2%.

Although students who enter the PharmD program have completed pre-requisites, they still have a diverse educational background and a range of experiences with science, especially the application of science in the context of genuine scientific research. At the COP only 31% of the students have completed a bachelor’s degree prior to entering the program.

Third year students were selected because of their curricular stage. In the second year of the curriculum the students were introduced to pharmacy research and informatics, which was their first formal exposure to literature evaluation in the pharmacy curriculum. In their fourth year, when they begin rotations, students are often required to use the medical literature to address a clinical question (as might occur in a clinical setting post-graduation). Placing this work in the third year adds to what students have learned in their second year and prepares them for the work expected in their last year prior to graduation.

The third year students who participated in the study completed the first clinical literature evaluation activity (implemented by this author) during their second year pharmacy informatics and research class. Thus, they have been exposed to the CLE rubric.

**Major Eligibility Criteria**

All third year Doctor of Pharmacy students were invited to participate in the study. Students who opted out of the study did not complete the two surveys but were
still required to complete all activity class work. That work was not included in the analysis.

**Research Procedures Involving Participants**

This research involved normal educational practices and, because of the nature of the educational activity (evaluating medical literature), was very low risk. The research fell under exempt category 2 of the Human Research Protection definitions because it involved surveys, instructional observations, and participants had the potential to be identified.

Students were un-identifiable except by the lab administrative staff who randomly assigned students a number to remove student identities while allowing matching student’s surveys and lesson work for data analysis. Students used their assigned number on all work and internet-based surveys. The administrative staff used the randomly assigned numbers to give students credit/no credit for completing their work for class grade purposes.

The study was presented to the students on the first day of class in the Spring 2012 semester and they were asked to provide or deny consent using a consent form. Analysis of the student work for purposes of the study differs from the reviews the instructor conducted for class grades. Student work grades and analysis of student work for the study remained separate and unshared between the instructor and myself.

I made classroom observations while the activity was being implemented for both lesson design considerations and data analysis considerations (e.g. differences in delivery between sections). As part of the observations, I listened in on groups of students, only
for purposes of monitoring the flow/implementation of the activity. No students were interviewed or identified during observations.

**Study Procedures**

The lesson was conducted over a period of one week and one day in the Spring 2012 semester in the third year lab using in-class sessions (small and large group works/discussion and individual on-line work) and out-of-class reading and online activities. The lab session contained 29-30 students per class and consisted of a single session 3-hours per week, repeated over 3 days so that all 89 students received the lesson.

The study procedures, and the associated approaches to data analysis, are outlined below. All data analysis occurred at the end of the activity to minimize bias. All student work was blinded to the investigator. Some work was collected online using Opineo survey software, some was collected electronically in Word via a special email account accessed only by the lab administrative assistant, and other work was paper versions.

1. In the week prior to the activity, students were introduced to the lesson and verbally informed of the study during the large class section (which has all 89 students). The following information was reviewed:

   - data are being collected to edit/improve the activity for future use in third year lab classes and for dissertation research;
   - the modified lessons and associated data will be shared with the pharmacy education and science education community via publications to advance the pharmacy EBM and science education fields;
   - student work is de-identified from the study investigator by having an assigned number that allows for matching of student work;
• all students will be expected to complete the work as part of their lab grade. They will receive credit-no credit from the instructor for completing each element. The analysis of their work for the research is not part of their grade and will not be shared with the instructor;

• students will be given the opportunity to provide direct feedback about the lesson in a survey;

2. During the week prior to implementing the lesson in class, students completed a brief, online, non-validated characterization survey (developed previously for the pilot study for this work) measuring their perceptions of confidence in and experience with reading, evaluating, understanding and applying results from a variety of medical studies on the same topic towards a clinical conclusion. The characterization survey is included in Appendix B.

   a. Survey data was analyzed using basic descriptive statistics including averages and percentages in Excel.

3. In addition, prior to the lab class (which was the week of January 23, 2012), students were instructed to read the cohort study (Kruetz, et al., 2010) and case-control (Jurrlink, et al., 2009) study. Students did not have access to the RCT at this time. They were instructed to evaluate the cohort and case control studies using the MLE rubric, identify four crux vocabulary words from the studies, rank the two studies relative to each other, and provide a preliminary clinical conclusion based upon those two studies, again, prior to class. Students were asked the following question after reading the cohort and case control studies: “What would your clinical conclusion be related to this topic if you just based
your decision upon the cohort and case control articles?" Because of the amount of pre-class work expected of the students, the large-class lecture (all 89 students) for that week was cancelled to provide additional time for students to do the work carefully. Students were expected to attend the smaller group lab session (28-29 students), which was when other elements of the activity occurred. Student instructions and worksheets are provided in Appendices C and D, the MLE Rubric is in Appendix E:

a. Students scored the articles using the MLE Rubric (in Word) and completed the crux vocabulary (also in Word) to capture the data electronically (they were emailed). Students only used their assigned number as an identifier on the document. The administrative assistant retrieved the student work from an email address created only for this study, gave credit to the students for their grade, de-identified any students who forgot to use their number, and then organized the data for the PI, ensuring the only identifiers present on the documents were the assigned numbers. The article ranking and initial clinical conclusion were completed online using the Opinio survey instrument. The survey was set up with complete anonymity parameters and had a field where students entered their assigned number.

b. The pre-lab work served as a measure for two elements: (1) student skill and knowledge level coming into the activity (these students did complete the single RCT activity in their second year informatics and research class); and (2) the extent that the MLE Rubric and crux vocabulary can
provide “guidance” to the students in evaluating the medical literature with their existing background knowledge, independent of the class discussion scaffolds (a few students worked together for the pre-lab, but the end-of-activity survey showed the majority worked independently).

c. Data analysis descriptions are provided in #5 below.

4. The in-lab elements of the activity and study are described next. For the first hour of class, the instructor (one of the experts in this study) reviewed the first two articles (cohort and case control studies), highlighting some of the key expert MLE Rubric scores and why those scores were given. He also reviewed the crux vocabulary identified by the experts. Next, the students had 1.5 hours to independently read the third article (the RCT) (Bhatt et al., 2010), individually complete the MLE Rubric for the RCT, talk about the RCT in small peer group discussions, and then individually reach his or her own clinical conclusion based on all three articles. This served as a measure of the contribution to student learning from the instructor/expert reviewing what the students had completed and from his providing the expert perspective on the cohort and case control studies (i.e. could the students transfer this knowledge from the other articles to a new article, the RCT?) and the small peer group discussion about the RCT (which did not have expert input at that point). The activity concluded by having the students work again in small groups of 3-4 to discuss all three of the articles and their clinical conclusions, followed by an instructor-lead whole-class discussion about the expert’s review of the RCT and the expert’s clinical conclusion. No student work data were collected after this, because it might have been more of a
measure of the students repeating what the expert’s said, rather than students’ own independent conclusions. A disadvantage of not collecting data at this point is not measuring potential contributions of these scaffolds to student learning.

a. The MLE Rubric data were analyzed quantitatively and qualitatively. Quantitative analysis consisted of comparing students’ MLE Rubric scores (total and section-by-section) with the expert consensus score for each part. Rubric data for all three articles were analyzed after the activity was completed using basic descriptive statistics including score distributions, averages, and percentages. Basic statistical calculations for this analysis were done using Excel.

b. Crux vocabulary words were analyzed qualitatively comparing which words students selected as crux vocabulary and their reasoning with that of the experts. Responses were categorically sorted into “weak,” “definition,” “partial,” and “reason.” The criteria for categorization are defined in Chapter 4.

c. Correlation analyses were also conducted to determine if there were any relationships between a student’s ability to identify and properly define/describe a crux vocabulary word and the MLE Rubric Scores. The analysis was done on SPSS for Mac OSX version 21.0.0

d. After reading the third article, the RCT, students were asked “What would your clinical conclusion be related to this topic if you just based your decision upon the three articles?” The student’s clinical conclusions were compared to the expert’s clinical conclusion. This analysis identifies the
extent to which the third article (the RCT) and the instructor/expert discussion of the cohort and case control studies scaffold impacted the student decisions. The final student clinical conclusions (which considers all three articles) was also compared to the expert’s clinical conclusion. An online z-test for proportions for dependent groups was run to determine if a statistically significant number of students changed their clinical conclusion (http://www.surveystar.com/our_services/ztest.htm).

5. At the end of the lesson students completed a survey (Appendix G) asking them to rate their perception of each scaffold’s contribution to their learning.
   a. Data were analyzed using basic descriptive statistics including averages and percentages.

Risks and Discomforts

Other than being a long, challenging assignment, there were no known risks or discomforts. Students were expected to complete their work in a normal fashion primarily with in-class time but with some homework. Analysis of the student work for the study was separate from the credit/no credit grade they received for their work.

Confidentiality and Privacy

Students were de-identified except to the administrative assistant, who informed the instructor which students should receive credit/no credit their work. Analysis of individual student work for the research was not shared with instructor (she will be able to see de-identified summary data if requested). Classroom observations were oriented towards the delivery of the lesson and students were not identified during in-class observations (the PI does not know the students).
Methods for Obtaining Expert MLE Scores, Article Rankings, and Clinical Conclusions

Three cardiac pharmacist-experts were assembled in December 2011 for a total of four hours to generate a consensus score for each of the three medical literature articles associated with the activity. Prior to the meeting each expert read and scored the articles using the MLE rubric and then in the meeting discussed each sub-section of the rubric for each article (e.g. Background section for the RCT, cohort and case control, rather than going through the rubric in its entirety for each article individually) reaching a single consensus score.

I recorded the consensus MLE rubric score using an Excel spread sheet and took notes about their reasoning for the score both within Excel and Word. Following the group meeting, each of the experts emailed their clinical conclusion based upon the articles. The clinical conclusions were similar and combined into a single conclusion representing the “expert clinical conclusion.” The experts’ consensus scores and clinical conclusions are presented in the results.

Multiple changes were made to the MLE rubric as a result of input from the experts as they went through the process. Changes consisted of: (1) word-smithing to provide better distinction between scoring levels or for clarification; (2) placing limits on scoring for specific study types; and (3) adding concepts/sub-sections to the rubric. The edited version of the MLE Rubric was used in this study.

Selection of the Medical Articles

Three medical research publications on the same topic were selected for the activity (Juurlink et al. 2009; Kruetz et al., 2010; Bhatt et al., 2010). The topic is a current
controversy investigating whether certain proton pump inhibitors (PPIs) interfere with the anti-blood clotting drug clopidogrel resulting in an increased risk of major adverse cardiovascular events such as myocardial infarction (MI). PPIs are often co-prescribed with clopidogrel because clopidogrel has been shown to increase GI bleeding. This potential drug interaction is an area of major concern because of the very large number of co-concomitant prescriptions given of these drugs across the world.

The three articles that were selected for the activity represent the three major types of large-scale study designs used in medical research – the randomized control trial, the cohort and the case-control. Differences between the types of studies are:

- **Randomized Control Trial (RCT):** An RCT is a form of a clinical trial where subjects in a population are randomly assigned to receive or not receive a treatment/exposure/intervention. Both groups are followed-up equally and measured for specific outcomes. RCTs are the “gold standard” study design.

- **Cohort Study:** Cohort studies are population-based studies that compare a group of people with an exposure to a different group (no exposure or different level of exposure) to answer the question “What are the effects of this exposure?” or “What are the risk factors associated with this exposure?” The studies may be prospective (“where the exposure is defined and subjects are selected before the outcome occurs”) or retrospective (“where the exposure is assessed after the outcome is known, usually by examination of medical records.”) (Scottish Intercollegiate Guidelines, 2004). Cohort studies need to observe enough subjects for a long-enough time to generate reliable incident and mortality rates (they are usually large numbers and
long-term) (Liddle, Williamson and Irwig, 1996). Cohort studies are good for identifying incidence and natural history of a disease and “can examine multiple outcomes after a single exposure” (Grimes and Schulz, 2002). It is the best way to assess absolute risk (Liddle, Williamson and Irwig, 1996).

- Case-Control Study (CC): A case-control study is an observational study comparing cases with a specific disease with a group without the disease ("controls"). These studies investigate an association between the hypothesized exposure and the disease being studied trying to answer the question “What exposures/conditions resulted in the disease?” Case-control studies are designed to estimate odds/relative risk. Ideally controls need to come from the same population as the cases to “reduce the chance that some other difference between the groups is accounting for the difference in the exposure that is under investigation.” Because case-control studies depend upon retrospective data, primarily in the form of people remembering their exposures, recall bias is possible. The use of biologic markers reduce the problem of recall bias (retrieved from http://www.gwumc.edu/library/tutorials/studydesign101/casecontrols.html).
Chapter 4

Results

The goal of this study was to test a scaffolded instructional model in a complex Evidence Based Medicine (EBM) lesson where students evaluated three different medical studies (RCT, case-control, and cohort) on the same topic and used those studies to reach an evidence-based conclusion. The hypothesis was that the students’ literature evaluation skills and subsequent application of the literature to address a clinical question would more closely approximate the experts following implementation of the model. The results do not fully support the hypothesis.

The results are presented in the following sections: Consents, Characterization Survey, MLE Rubric Scores, Article Ranking, Crux Vocabulary, Clinical Conclusion and Scaffold Contributions Survey.

Consents

Of the 89 students in the class 56 consented to participate in the study and 33 said no. Students potentially said no because of the surveys to complete on top of the activity, which take additional time. Not all of the consenting students completed all elements of the study. Each number of students in the sample is defined with each data set.

Characterization Survey

The characterization survey is an unvalidated survey that was originally developed for the pilot study and used for this research. It attempted to measure student perceptions of evidence-based medicine, and their confidence in and experience with reading, evaluating, understanding and applying results from a variety of medical studies on the same topic towards a clinical conclusion (Appendix B). The survey helped
characterize the background level of student experience prior to this activity, as well as, to provide a tool for comparing student self-perceptions with their work level relative to experts.

Forty-nine (49) students completed the characterization survey. The results are presented in Tables 3-8. When asked how familiar students were with evidence-based medicine the majority (49%) indicated they were moderately familiar with it, and 35% selected the option of “somewhat familiar.” Only 16% selected “very familiar.” Despite the relatively lower familiarity, 80% indicated that evidence-based medicine was “very important” in pharmacy practice (Table 2).

Table 2

*Characterization Survey Results: Familiarity With and Importance of EBM*

<table>
<thead>
<tr>
<th></th>
<th>How familiar are you with evidence-based medicine?</th>
<th>How important should evidence-based medicine be in pharmacy practice?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not at All</td>
<td>0% (0)</td>
<td>0% (0)</td>
</tr>
<tr>
<td>Somewhat</td>
<td>35% (17)</td>
<td>0%</td>
</tr>
<tr>
<td>Moderately</td>
<td>49% (24)</td>
<td>21% (10)</td>
</tr>
<tr>
<td>Very</td>
<td>16% (8)</td>
<td>80% (39)</td>
</tr>
</tbody>
</table>

(n=49)

The next set of questions inquired about student experience conducting research (Table 3), reading and evaluating primary scientific literature (i.e. full publications of scientific research from any scientific journal) and medical literature (Table 4), and publishing research themselves (Table 5). Although 49% of the students have participated in some form of scientific research, 80% of them have never been an author or co-author on a paper.
Table 3

*Characterization Survey Results: Student Participation in scientific Research*

<table>
<thead>
<tr>
<th>Have you personally participated in conducting scientific research? If yes, select all that apply.</th>
<th>Response Frequency (% represent relative frequency by choice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>52% (29)</td>
</tr>
<tr>
<td>Yes, as a high school student</td>
<td>2% (1)</td>
</tr>
<tr>
<td>Yes, as a pre-pharmacy student</td>
<td>21% (12)</td>
</tr>
<tr>
<td>Yes, as a pharmacy student</td>
<td>16% (9)</td>
</tr>
<tr>
<td>Yes, in a previous job</td>
<td>9% (5)</td>
</tr>
</tbody>
</table>

(n=56, multiple responses allowed)

Table 4

*Characterization Survey Results: Number of Scientific Literature Publications Read by Students*

<table>
<thead>
<tr>
<th>Approximately how many primary scientific literature publications have you read (this can include any scientific or medical subject)</th>
<th>Response Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fewer than 5</td>
<td>10% (5)</td>
</tr>
<tr>
<td>5-10</td>
<td>10% (5)</td>
</tr>
<tr>
<td>11-20</td>
<td>27% (13)</td>
</tr>
<tr>
<td>21-40</td>
<td>39% (19)</td>
</tr>
<tr>
<td>&gt;40</td>
<td>14% (7)</td>
</tr>
</tbody>
</table>

(n=49)

Table 5

*Characterization Survey Results: Percent of Students Co-Authoring A Scientific Paper*

<table>
<thead>
<tr>
<th>Have you been an author or co-author on a scientific paper?</th>
<th>Response Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>80% (39)</td>
</tr>
<tr>
<td>Yes</td>
<td>20% (10)</td>
</tr>
</tbody>
</table>

(n=49)

When asked how often they read primary scientific literature (Table 6) 43% indicated a few times per month. Thirty-one percent indicated a few times per year. They were also asked to estimate the number of scientific papers they have read to date, 39%
selected 21-40 papers. The next most frequent selection was 11-20 papers (27%). Fifteen percent (15%) of students said they have read more than 40 scientific papers, while the remaining 20% indicated they have read fewer than 10.

Table 6

Characterization Survey Results: Frequency Reading Scientific Literature

<table>
<thead>
<tr>
<th>How often do you read primary scientific literature (i.e. full publications of the original research) from research journals (this includes medical and bench research)?</th>
<th>Response Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not at all</td>
<td>2% (1)</td>
</tr>
<tr>
<td>Rarely</td>
<td>18% (9)</td>
</tr>
<tr>
<td>A few times per year</td>
<td>31% (15)</td>
</tr>
<tr>
<td>A few times per month</td>
<td>43% (21)</td>
</tr>
<tr>
<td>A few times per week</td>
<td>6% (3)</td>
</tr>
</tbody>
</table>

(n=49)

Students were presented questions about their confidence in reading and understanding a research study, as well as their confidence in being able to evaluate the quality of a study (Table 7). Responses associated with “reading and understanding” a research publication had a comparatively higher percentage of “mostly” responses, indicating higher confidence, compared to the evaluation questions. For example, 76% of the respondents indicated “mostly” when asked they could “confidently read and understand a published research study in (their) field of interest.” The percentage of students who felt confident reading and understanding a published research study outside of their field of interest dropped with 49% selecting “a little” for the statement and 41% selecting “mostly.” Fifty-three percent indicated that they were mostly confident in their “ability to look up or obtain information to help (their) understanding of a published research study.”
The next series of questions attempted to identify student perceptions of confidence in evaluating specific study types (RCT, cohort, or case control). The responses were similar for all three study types with greater confidence in RCT and cohort. When asked about their confidence evaluating a case-control study, 49% indicated “a little” and 47% indicated “mostly.” When presented with the statement “I can confidently evaluate a randomized control trial,” 49% selected “mostly” and 41% selected “a little.” When asked about their confidence in evaluating a cohort study, students were nearly evenly split, with 47% indicating “a little” and 49% indicating “mostly.” Eight percent (8%) felt they “definitely” could read and understand a study outside of their field. Trends were similar for the other study types but with more students trending towards greater confidence.

There was a similar split with the question “I know what specific things to look for to determine the quality of a single, published research study.” Only 1 student selected “not at all,” and 43% and 51% selected “a little” and “mostly” respectively. Two students (4%) selected “definitely.” When asked whether they know specific things to look for to determine the quality of a research journal 8% indicated “not at all,” 41% indicated “a little,” and 49% indicated “mostly.” Only 1 student selected the highest option of definitely.

Interestingly, their self-ratings increased when asked to rate their confidence in making a “clinical conclusion using multiple individual studies (excluding reviews and clinical guidelines)” with 59% of the respondents indicating “mostly” and 6% indicating “definitely.” Only one respondent (2%) indicated “not at all” and the remaining 33% selected “a little.” This result is interesting because it seems to indicate a potential
disconnect between the role of evaluating the quality of literature and reaching a clinical conclusion.

Table 7

*Characterization Survey Results: Confidence Questions*

<table>
<thead>
<tr>
<th></th>
<th>Not at All</th>
<th>A Little</th>
<th>Mostly</th>
<th>Definitely</th>
</tr>
</thead>
<tbody>
<tr>
<td>I know what specific things to look for to determine the quality of a research journal</td>
<td>8% (4)</td>
<td>41% (20)</td>
<td>49% (24)</td>
<td>2% (1)</td>
</tr>
<tr>
<td>I know what specific things to look for to determine the quality of a single, published study</td>
<td>2% (1)</td>
<td>43% (21)</td>
<td>51% (25)</td>
<td>4% (2)</td>
</tr>
<tr>
<td>I am confident in my ability to look up or obtain information to help my understanding of a published research study</td>
<td>2% (1)</td>
<td>35% (17)</td>
<td>53% (26)</td>
<td>10% (5)</td>
</tr>
<tr>
<td>I can confidently read and understand a published research study inside of my field of interest</td>
<td>2% (1)</td>
<td>14% (7)</td>
<td>76% (37)</td>
<td>8% (4)</td>
</tr>
<tr>
<td>I can confidently read and understand a published research study outside of my field of interest</td>
<td>6% (3)</td>
<td>49% (24)</td>
<td>40% (20)</td>
<td>4% (2)</td>
</tr>
<tr>
<td>I can confidently evaluate a randomized control trial (RCT)</td>
<td>2% (1)</td>
<td>41% (20)</td>
<td>49% (24)</td>
<td>8% (4)</td>
</tr>
<tr>
<td>I can confidently evaluate a cohort study</td>
<td>2% (1)</td>
<td>47% (23)</td>
<td>49% (24)</td>
<td>2% (1)</td>
</tr>
<tr>
<td>I can confidently evaluate a case-control study</td>
<td>2% (1)</td>
<td>49% (24)</td>
<td>47% (23)</td>
<td>2% (1)</td>
</tr>
<tr>
<td>I can confidently make a clinical decision using multiple individual studies (excludes reviews and guidelines)</td>
<td>2% (1)</td>
<td>33% (16)</td>
<td>59% (29)</td>
<td>6% (3)</td>
</tr>
</tbody>
</table>

(n= 49)
The last question of the characterization survey asked “What do you do if individual study results conflict and there are no reviews or clinical guidelines to which to refer?” There were 44 open-ended responses to the question which were sorted into the categories “vague,” “defer to expert advice,” “some reference to using literature but not evaluating literature,” “application of literature evaluation” and categories combined with “defer to expert advice” (Table 8). The response that was sought or considered “ideal” for this question was one that included a reference to evaluating the quality of the studies and giving a greater weight to the higher quality study in making a clinical conclusion. An example of a student response from each category is provided below:

- Vague – “Err on the side of caution and take a more conservative approach.”
- Defer to expert advice – “I would call a colleague to get their opinion on what to do.”
- Some reference to using literature but not evaluating literature – “Search for other published research studies that pertain to the topic.”
- Application of literature evaluation (“ideal response category”) - “Careful analysis of the data presented and how it was obtained. This includes the type of study, statistical methods employed, where the information was obtained from, author credentials (if available), bias, etc.”

Only 34% of the responses articulated some application of literature evaluation and relative weighting of the articles to help guide the clinical conclusion. Twenty percent of the responses solely mentioned consulting with/deferring to experts in the field and 14% made some reference to using medical literature in their decision-making. The
remaining comments were either vague or fell into the “other” category. When looking at the total responses, 64% did not refer to evaluating and weighing the quality of the articles to reach a clinical conclusion.

Table 8

Characterization Survey Results: How Student Would Address Conflicting Study Results

<table>
<thead>
<tr>
<th>Application of literature evaluation</th>
<th>Some reference to using literature but does not mention evaluation</th>
<th>Defer to or rely on expert advice</th>
<th>Other</th>
<th>Vague</th>
</tr>
</thead>
<tbody>
<tr>
<td>36% (16)</td>
<td>14% (6)</td>
<td>20% (9) only refer to expert advice</td>
<td>2% (1)</td>
<td>11% (5)</td>
</tr>
<tr>
<td>SUB-TOTAL COMBINED 52%</td>
<td>18% (7) of the total responses overlap between the above two categories</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL COMMENTS THAT DO NOT MENTION LITERATURE EVALUATION</td>
<td>64% (28)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These responses indicate limited orientation towards EBM, which is consistent with their indication of moderate familiarity with EBM, despite students’ indications that EBM was important.

MLE Rubric Scores, Article Ranking, Crux Vocabulary, and Initial Clinical

Conclusion

As part of the pre-lab activity, students were asked to read and score two of the three studies (the cohort and case control) using the MLE rubric and then reach an initial
clinical conclusion based upon those two studies. This provides a measure of the student’s relative background knowledge prior to the lab scaffold inputs and the utility of the MLE Rubric as a scaffold by itself to guide students in article evaluation.

It also provides a measure of the clinical conclusion students would make when based solely upon the two articles. In the lab, one of the experts reviewed MLE rubric highlighting the strengths and weaknesses of the cohort and case control studies (scaffold) according to the experts. Following the in-class review, students were asked to read the RCT study, talk in small groups (scaffold) and then individually score the RCT. The RCT MLE Rubric scores provide a measure of student ability to evaluate the study following the implementation of two additional scaffolds (an instructor-lead review of key points related to evaluating the cohort and case control studies and the small group discussions).

Prior to coming to class, students were also asked to identify and define/describe four crux vocabulary words from the cohort and case control studies (scaffold). The definition of crux vocabulary is “words or concepts that must be understood in-depth to be able to evaluate the quality of the study.” Students were provided a detailed example of a crux vocabulary word (Appendix D). Although students have been asked to identify vocabulary words in a previous medical literature evaluation activity conducted in the beginning of their second year, the concept of crux vocabulary is new, but potentially important in differentiating a student’s level of understanding of the complex interaction between medical study designs and pharmacy content knowledge. The data from these assessments are described in more detail in the following sub-sections: MLE Rubric Scores; Article Ranking; Crux Vocabulary; and Initial Clinical Conclusion.
**MLE Rubric scores.** Students were given a copy of the MLE rubric (Appendix E) and asked to score the cohort and case-control studies prior to coming to class. These data serve as a measure of the utility of the MLE Rubric as an independent scaffold in instruction and, when compared to the experts, provide a reference point for student knowledge of evaluating medical literature in the third year of the degree program. (The student article scoring is one of the most important and data-rich aspects of this research as it reveals where they are in their learning and understanding of critically evaluating medical literature compared to expert-level analysis.)

The rubric, through its detailed descriptions and blacked out sections, is intended to help train the students on the many elements to consider when evaluating medical literature. Some sections of the rubric are blacked-out because those items are not relevant for a specific study type. For example, “participation rate” and “randomization” consistently do not apply to cohort or case control studies.

Other medical study elements are not blacked-out but may be “not applicable” (N/A) in certain circumstances. For example, “blinding,” “interim analyses,” and “participant flow” are not necessary for all studies. However, being able to discern whether it applies to a given study requires a certain level of sophisticated knowledge. Thus, in circumstances where N/A should be applied, students must be able to judge whether or not to assign a score from 0-3 or to apply N/A. When N/A is given, they must then subtract that item from the denominator to calculate the average score for the study.

The option of applying N/A in certain circumstances demonstrates the complexity of evaluating medical literature, but it also complicates the data analysis. For example, there are places where students should have assigned N/A and did not (they gave a score
Instead), and vice-versa, where they assigned N/A and should not. In multiple circumstances, N/A was not even an option but numerous students still indicated N/A.

Incorrectly applying N/A is important to consider in the results, however assumptions sometimes need to be made about what students intended. The default assumption for this research is that the student wrote down exactly what s/he meant. For example, there are numerous circumstances where students placed a 0 when it should have been an N/A. The natural question is did they intend to assign a score of 0 or did they intend to put N/A? In terms of the rubric, a 0 is a real score with specific descriptors. A 0 means the information should have been there but it was not provided or done by the authors. Whereas, an N/A means that it was not applicable to that study design and should not have been included in any descriptions (which is also why an N/A is removed from the denominator). If a student placed a 0 instead of an N/A, it is counted as an intended score (which increases the denominator and impacts the total percent score assigned by the student to the study).

The data analysis assumption that a 0 was intended by a student to be a 0 rather than an N/A is supported in two ways. One, as mentioned above, 0 and N/A are defined in the rubric. Second, many students who placed a 0 instead of an N/A used N/A in some other location in their scoring (whether or not it was a correct use of N/A) thereby indicating they differentiated between 0 and N/A.

The data presented in this section (Tables 9-11) are broken down by study type and include:

- The expert consensus score;
• Average and median MLE Rubric Scores for each section of the rubric and total percent score for each study type (cohort, case control, RCT) compared to the expert consensus score;

• The range of student MLE Rubric Scores for each section of the MLE Rubric;

• The percent of students who gave the same score as the expert for each section of the MLE Rubric; and

• The number of N/As that were incorrectly assigned to a given category;

Tables 9-11 provide comprehensive presentation of the data by study type. Certain fields are highlighted to emphasize the differences between the student and expert scores (if there is greater than or equal to a 0.7 point difference, if students assigning the same score as the experts fell below 60%, or if there were any incorrectly assigned N/As).
Table 9

Summary Data for Student Pre-Lab Cohort MLE Rubric Scores Compared to the Expert Consensus Score

<table>
<thead>
<tr>
<th>MLE Rubric Section</th>
<th>Expert Consensus Score</th>
<th>Average Student MLE Rubric Score</th>
<th>Median Student Scores</th>
<th>Range of Student MLE Rubric Scores</th>
<th>% of Students Assigning the Same Score as the Expert Consensus Score</th>
<th>Percent of N/As that were Incorrectly Assigned (number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>Background</td>
<td>3</td>
<td>2.6</td>
<td>3</td>
<td>1-3</td>
<td>64%</td>
</tr>
<tr>
<td></td>
<td>Objectives</td>
<td>3</td>
<td>2.7</td>
<td>3</td>
<td>1-3</td>
<td>77%</td>
</tr>
<tr>
<td>Study Design/Methods</td>
<td>Study Design</td>
<td>3</td>
<td>2.3</td>
<td>2</td>
<td>1-3</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>Blinding (N/A option)</td>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td>N/A, 0-3</td>
<td>59%</td>
</tr>
<tr>
<td></td>
<td>Control/Comparison Group</td>
<td>2</td>
<td>2.0</td>
<td>2</td>
<td>N/A, 0-3</td>
<td>64%</td>
</tr>
<tr>
<td></td>
<td>Eligibility Criteria</td>
<td>2</td>
<td>2.0</td>
<td>2</td>
<td>N/A, 0-3</td>
<td>57%</td>
</tr>
<tr>
<td></td>
<td>Generalizability</td>
<td>3</td>
<td>2.2</td>
<td>2</td>
<td>1-3</td>
<td>38%</td>
</tr>
<tr>
<td>Participation Rate</td>
<td>Randomization 1 (from Large Population)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Randomization 2 (Treatments)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Implementation (Enrollment/Concealment)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Treatment Plan/Intervention 1</td>
<td>2</td>
<td>2.2</td>
<td>2</td>
<td>N/A, 0-3</td>
<td>38%</td>
</tr>
<tr>
<td></td>
<td>Treatment Plan/</td>
<td>2</td>
<td>2.2</td>
<td>3</td>
<td>N/A, 0-3</td>
<td>32%</td>
</tr>
<tr>
<td>MLE Rubric Section</td>
<td>Expert Consensus Score</td>
<td>Average Student MLE Rubric Score</td>
<td>Median Student Scores</td>
<td>Range of Student MLE Rubric Scores</td>
<td>% of Students Assigning the Same Score as the Expert Consensus Score</td>
<td>Percent of N/As that were Incorrectly Assigned (number)</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>------------------------</td>
<td>---------------------------------</td>
<td>-----------------------</td>
<td>-----------------------------------</td>
<td>---------------------------------------------------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Intervention 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporality</td>
<td>3</td>
<td>2.1</td>
<td>2</td>
<td>0-3</td>
<td>32%</td>
<td>0%</td>
</tr>
<tr>
<td>Measure of Exposure Assessment</td>
<td>2</td>
<td>2.0</td>
<td>2</td>
<td>0-3</td>
<td>51%</td>
<td>0%</td>
</tr>
<tr>
<td>Changes in Study Design</td>
<td>3</td>
<td>2.5</td>
<td>3</td>
<td>N/A, 0-3</td>
<td>66%</td>
<td>7% (4)</td>
</tr>
<tr>
<td>Outcomes Described</td>
<td>3</td>
<td>2.5</td>
<td>3</td>
<td>2-3</td>
<td>54%</td>
<td>0%</td>
</tr>
<tr>
<td>Measures/ Methods Described/ Justified</td>
<td>2</td>
<td>2.1</td>
<td>2</td>
<td>1-3</td>
<td>66%</td>
<td>0%</td>
</tr>
<tr>
<td>Quality Improvement</td>
<td>1</td>
<td>1.4</td>
<td>1</td>
<td>0-3</td>
<td>29%</td>
<td>0%</td>
</tr>
<tr>
<td>Sample Size (Power Calculations)</td>
<td>3</td>
<td>2.4</td>
<td>3</td>
<td>0-3</td>
<td>63%</td>
<td>0%</td>
</tr>
<tr>
<td>Sample Size (Enrolled and Completed)</td>
<td>3</td>
<td>2.7</td>
<td>3</td>
<td>0-3</td>
<td>77%</td>
<td>0%</td>
</tr>
<tr>
<td>Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical Methods</td>
<td>3</td>
<td>2.4</td>
<td>2</td>
<td>1-3</td>
<td>45%</td>
<td>0%</td>
</tr>
<tr>
<td>Confounding Factors</td>
<td>2</td>
<td>1.8</td>
<td>2</td>
<td>0-3</td>
<td>61%</td>
<td>0%</td>
</tr>
<tr>
<td>Interim Analyses (N/A option)</td>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td>N/A, 0-3</td>
<td>86%</td>
<td>14% (8)</td>
</tr>
<tr>
<td>Results/ Conclusions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant Flow (N/A Option)</td>
<td>3</td>
<td>2.6</td>
<td>3</td>
<td>N/A, 0-3</td>
<td>54%</td>
<td>20% (11)</td>
</tr>
<tr>
<td>Key Data</td>
<td>2</td>
<td>3.1</td>
<td>3</td>
<td>1-3</td>
<td>45%</td>
<td>0%</td>
</tr>
<tr>
<td>MLE Rubric Section</td>
<td>Expert Consensus Score</td>
<td>Average Student MLE Rubric Score</td>
<td>Median Student Scores</td>
<td>Range of Student MLE Rubric Scores</td>
<td>% of Students Assigning the Same Score as the Expert Consensus Score</td>
<td>Percent of N/As that were Incorrectly Assigned (number)</td>
</tr>
<tr>
<td>----------------------------</td>
<td>------------------------</td>
<td>---------------------------------</td>
<td>----------------------</td>
<td>-----------------------------------</td>
<td>-------------------------------------------------------------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>Numbers Analyzed</td>
<td>3</td>
<td>2.3</td>
<td>2.5</td>
<td>1-3</td>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td>Outcomes and Estimation</td>
<td>3</td>
<td>2.9</td>
<td>3</td>
<td>2-3</td>
<td>88%</td>
<td>0%</td>
</tr>
<tr>
<td>Ancillary Analyses</td>
<td>3</td>
<td>1.7</td>
<td>2</td>
<td>0-3</td>
<td>20%</td>
<td>0%</td>
</tr>
<tr>
<td>Adverse Events/ Harms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias/ Conflict of Interest</td>
<td>3</td>
<td>2.1</td>
<td>3</td>
<td>0-3</td>
<td>55%</td>
<td>0%</td>
</tr>
<tr>
<td>Interpretation (Consistent with Data)</td>
<td>3</td>
<td>2.5</td>
<td>3</td>
<td>1-3</td>
<td>57%</td>
<td>0%</td>
</tr>
<tr>
<td>Overall Evidence of Study</td>
<td>3</td>
<td>2.3</td>
<td>2</td>
<td>1-3</td>
<td>41%</td>
<td>0%</td>
</tr>
<tr>
<td>Total % Score</td>
<td></td>
<td>87%</td>
<td>75%</td>
<td>53%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. (n= 56) – Fields are highlighted (and the corresponding category) if there is a ≥0.7 point difference between the expert and student scores, if the % if students assigning the same score as the experts falls below 60%, and if there were any incorrectly assigned N/As.
### Table 10

*Summary Data for Student Pre-Lab Case Control MLE Rubric Scores Compared to the Expert Consensus Score*

<table>
<thead>
<tr>
<th>MLE Rubric Section</th>
<th>Expert Consensus Score</th>
<th>Student Average MLE Rubric Score</th>
<th>Median Student Scores</th>
<th>Range of Student MLE Rubric Scores</th>
<th>% of Students Assigning the Same Score as the Expert Consensus Score</th>
<th>Percent of N/As that were Incorrectly Assigned (number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Background</td>
<td>3</td>
<td>2.6</td>
<td>3</td>
<td>1-3</td>
<td>59%</td>
</tr>
<tr>
<td></td>
<td>Objectives</td>
<td>1</td>
<td>2.4</td>
<td>3</td>
<td>1-3</td>
<td>9%</td>
</tr>
<tr>
<td>Study Design/Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Study Design</td>
<td>3</td>
<td>2.2</td>
<td>2</td>
<td>1-3</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>Blinding (N/A option)</td>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td>N/A, 0-3</td>
<td>59%</td>
</tr>
<tr>
<td></td>
<td>Control/Comparison Group</td>
<td>2</td>
<td>1.8</td>
<td>2</td>
<td>0-3</td>
<td>52%</td>
</tr>
<tr>
<td></td>
<td>Eligibility Criteria</td>
<td>2</td>
<td>1.9</td>
<td>2</td>
<td>1-3</td>
<td>54%</td>
</tr>
<tr>
<td></td>
<td>Generalizability</td>
<td>1</td>
<td>1.7</td>
<td>2</td>
<td>1-3</td>
<td>41%</td>
</tr>
<tr>
<td>Participation Rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomization 1 (from Large Population)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomization 2 (Treatments)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation (Enrollment/Concealment)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment Plan/Intervention 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment Plan/Intervention 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>MLE Rubric Section</th>
<th>Expert Consensus Score</th>
<th>Student Average MLE Rubric Score</th>
<th>Median Student Scores</th>
<th>Range of Student MLE Rubric Scores</th>
<th>% of Students Assigning the Same Score as the Expert Consensus Score</th>
<th>Percent of N/As that were Incorrectly Assigned (number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure of Exposure Assessment</td>
<td>2</td>
<td>1.9</td>
<td>2</td>
<td>0-3</td>
<td>60%</td>
<td>0%</td>
</tr>
<tr>
<td>Changes in Study Design</td>
<td>3</td>
<td>2.1</td>
<td>3</td>
<td>N/A, 0-3</td>
<td>55%</td>
<td>7% (4)</td>
</tr>
<tr>
<td>Outcomes Described</td>
<td>2</td>
<td>2.0</td>
<td>2</td>
<td>0-3</td>
<td>48%</td>
<td>0%</td>
</tr>
<tr>
<td>Measures/ Methods Described/ Justified</td>
<td>1</td>
<td>1.9</td>
<td>2</td>
<td>0-3</td>
<td>23%</td>
<td>0%</td>
</tr>
<tr>
<td>Quality Enhancement</td>
<td>1</td>
<td>1.3</td>
<td>1</td>
<td>0-3</td>
<td>41%</td>
<td>0%</td>
</tr>
<tr>
<td>Sample Size (Power Calculations)</td>
<td>0</td>
<td>1.0</td>
<td>0</td>
<td>0-3</td>
<td>52%</td>
<td>0%</td>
</tr>
<tr>
<td>Sample Size (Enrolled and Completed)</td>
<td>1</td>
<td>1.7</td>
<td>2</td>
<td>0-3</td>
<td>20%</td>
<td>2% (1)</td>
</tr>
<tr>
<td>Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical Methods</td>
<td>2</td>
<td>2.2</td>
<td>2</td>
<td>1-3</td>
<td>39%</td>
<td>0%</td>
</tr>
<tr>
<td>Confounding Factors</td>
<td>1</td>
<td>1.6</td>
<td>2</td>
<td>0-3</td>
<td>27%</td>
<td>0%</td>
</tr>
<tr>
<td>Interim Analyses (N/A option)</td>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td>N/A, 0-3</td>
<td>79%</td>
<td>21% (12)</td>
</tr>
<tr>
<td>Results/ 1.5Conclusions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant Flow (N/A option)</td>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td>N/A, 0-3</td>
<td>61%</td>
<td>41% (23)</td>
</tr>
<tr>
<td>Key Data</td>
<td>2</td>
<td>2.3</td>
<td>2</td>
<td>1-3</td>
<td>39%</td>
<td>0%</td>
</tr>
<tr>
<td>Numbers Analyzed</td>
<td>3</td>
<td>2.3</td>
<td>2</td>
<td>1-3</td>
<td>43%</td>
<td>0%</td>
</tr>
<tr>
<td>MLE Rubric Section</td>
<td>Expert Consensus Score</td>
<td>Student Average MLE Rubric Score</td>
<td>Median Student Scores</td>
<td>Range of Student MLE Rubric Scores</td>
<td>% of Students Assigning the Same Score as the Expert Consensus Score</td>
<td>Percent of N/As that were Incorrectly Assigned (number)</td>
</tr>
<tr>
<td>----------------------------</td>
<td>------------------------</td>
<td>---------------------------------</td>
<td>-----------------------</td>
<td>-----------------------------------</td>
<td>---------------------------------------------------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Outcomes and Estimation</td>
<td>2</td>
<td>2.5</td>
<td>3</td>
<td>1-3</td>
<td>25%</td>
<td>0%</td>
</tr>
<tr>
<td>Ancillary Analyses</td>
<td>3</td>
<td>1.3</td>
<td>1.5</td>
<td>0-3</td>
<td>13%</td>
<td>0%</td>
</tr>
<tr>
<td>Adverse Events/ Harms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias/ Conflict of Interest</td>
<td>3</td>
<td>2.3</td>
<td>3</td>
<td>0-3</td>
<td>57%</td>
<td>0%</td>
</tr>
<tr>
<td>Interpretation (Consistent with Data)</td>
<td>3</td>
<td>2.3</td>
<td>2</td>
<td>0-3</td>
<td>45%</td>
<td>0%</td>
</tr>
<tr>
<td>Overall Evidence of Study</td>
<td>2</td>
<td>2.1</td>
<td>2</td>
<td>0-3</td>
<td>52%</td>
<td>0%</td>
</tr>
<tr>
<td>Total % Score</td>
<td>65%</td>
<td>71%</td>
<td></td>
<td></td>
<td>43%</td>
<td></td>
</tr>
</tbody>
</table>

*Note. (n= 56) – Fields are highlighted (and the corresponding category) if there is a ≥0.7 point difference between the expert and student scores, if the % if students assigning the same score as the experts falls below 60%, and if there were any incorrectly assigned N/As.*
Table 11

**Summary Data for Student Pre-Lab RCT MLE Rubric Scores Compared to the Expert Consensus Score**

<table>
<thead>
<tr>
<th>MLE Rubric Section</th>
<th>Expert Consensus Score</th>
<th>Average Student MLE Rubric Score</th>
<th>Median Student Scores</th>
<th>Range of Student MLE Rubric Scores</th>
<th>% of Students Assigning the Same Score as the Expert Consensus Score</th>
<th>% of N/As that were Incorrectly Assigned (Number)</th>
<th>% of Blank Responses (Number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>Background</td>
<td>1</td>
<td>2.2</td>
<td>2.0</td>
<td>0-3</td>
<td>11%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Objectives</td>
<td>1</td>
<td>2.6</td>
<td>3.0</td>
<td>1-3</td>
<td>8%</td>
<td>0%</td>
</tr>
<tr>
<td>Study Design/Methods</td>
<td>Study Design</td>
<td>3</td>
<td>2.5</td>
<td>3.0</td>
<td>1-3</td>
<td>53%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Blinding (N/A option)</td>
<td>2</td>
<td>2.4</td>
<td>2.0</td>
<td>0-3</td>
<td>49%</td>
<td>2% (1)</td>
</tr>
<tr>
<td></td>
<td>Control/Comparison Group</td>
<td>3</td>
<td>2.6</td>
<td>3.0</td>
<td>0-3</td>
<td>68%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Eligibility Criteria</td>
<td>3</td>
<td>2.3</td>
<td>2.0</td>
<td>1-3</td>
<td>40%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Generalizability</td>
<td>2</td>
<td>2.2</td>
<td>2.0</td>
<td>1-3</td>
<td>49%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Participation Rate</td>
<td>3</td>
<td>2.3</td>
<td>2.0</td>
<td>N/A, 0-3</td>
<td>38%</td>
<td>2% (1)</td>
</tr>
<tr>
<td></td>
<td>Randomization 1 (from Large Population)</td>
<td>3</td>
<td>2.5</td>
<td>3.0</td>
<td>0-3</td>
<td>62%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Randomization 2 (Treatments)</td>
<td>3</td>
<td>2.5</td>
<td>3.0</td>
<td>1-3</td>
<td>55%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Implementation (Enrollment/Concealment)</td>
<td>3</td>
<td>2.5</td>
<td>3.0</td>
<td>0-3</td>
<td>55%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Treatment Plan/Intervention 1</td>
<td>2</td>
<td>2.6</td>
<td>3.0</td>
<td>1-3</td>
<td>32%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Treatment Plan/</td>
<td>2</td>
<td>2.5</td>
<td>3.0</td>
<td>N/A, 1-3</td>
<td>30%</td>
<td>2% (1)</td>
</tr>
</tbody>
</table>
## Randomized Control Trial (RCT)

<table>
<thead>
<tr>
<th>MLE Rubric Section</th>
<th>Expert Consensus Score</th>
<th>Average Student MLE Rubric Score</th>
<th>Median Student Scores</th>
<th>Range of Student MLE Rubric Scores</th>
<th>% of Students Assigning the Same Score as the Expert Consensus Score</th>
<th>% of N/As that were Incorrectly Assigned (Number)</th>
<th>% of Blank Responses (Number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporality</td>
<td>2</td>
<td>1.8</td>
<td>2.0</td>
<td>0-3</td>
<td>21%</td>
<td>0%</td>
<td>2% (1)</td>
</tr>
<tr>
<td>Measure of Exposure Assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Changes in Study Design</td>
<td>1</td>
<td>1.9</td>
<td>2.0</td>
<td>N/A, 0-3</td>
<td>23%</td>
<td>4% (2)</td>
<td>8% (4)</td>
</tr>
<tr>
<td>Outcomes Described</td>
<td>3</td>
<td>2.6</td>
<td>3.0</td>
<td>1-3</td>
<td>64%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Measures/Methods Described/ Justified</td>
<td>2</td>
<td>2.4</td>
<td>2.0</td>
<td>0-3</td>
<td>47%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Quality Enhancement</td>
<td>3</td>
<td>1.9</td>
<td>2.0</td>
<td>N/A, 0-3</td>
<td>23%</td>
<td>4% (2)</td>
<td>2% (1)</td>
</tr>
<tr>
<td>Sample Size (Power Calculations)</td>
<td>0</td>
<td>1.9</td>
<td>2.0</td>
<td>0-3</td>
<td>11%</td>
<td>0%</td>
<td>2% (1)</td>
</tr>
<tr>
<td>Sample Size (Enrolled and Completed)</td>
<td>1</td>
<td>1.8</td>
<td>2.0</td>
<td>0-3</td>
<td>30%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical Methods</td>
<td>3</td>
<td>2.5</td>
<td>3.0</td>
<td>1-3</td>
<td>58%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Confounding Factors</td>
<td>1</td>
<td>2.0</td>
<td>2.0</td>
<td>0-3</td>
<td>21%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Interim Analyses</td>
<td>0</td>
<td>1.8</td>
<td>2.0</td>
<td>N/A, 0-3</td>
<td>8%</td>
<td>47% (25)</td>
<td>2% (1)</td>
</tr>
</tbody>
</table>
## Randomized Control Trial (RCT)

<table>
<thead>
<tr>
<th>MLE Rubric Section</th>
<th>Expert Consensus Score</th>
<th>Average Student MLE Rubric Score</th>
<th>Median Student Scores</th>
<th>Range of Student MLE Rubric Scores</th>
<th>% of Students Assigning the Same Score as the Expert Consensus Score</th>
<th>% of N/As that were Incorrectly Assigned (Number)</th>
<th>% of Blank Responses (Number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N/A option)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results/ 2Conclusions</td>
<td>Participant Flow (N/A Option)</td>
<td>3</td>
<td>1.7</td>
<td>2.0</td>
<td>N/A, 0-3</td>
<td>34%</td>
<td>4% (2)</td>
</tr>
<tr>
<td></td>
<td>Key Data</td>
<td>3</td>
<td>2.4</td>
<td>2.0</td>
<td>1-3</td>
<td>43%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Numbers Analyzed</td>
<td>3</td>
<td>2.5</td>
<td>3.0</td>
<td>1-3</td>
<td>58%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Outcomes and Estimation</td>
<td>3</td>
<td>2.7</td>
<td>3.0</td>
<td>1-3</td>
<td>70%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Ancillary Analyses</td>
<td>2</td>
<td>2.0</td>
<td>2.0</td>
<td>0-3</td>
<td>45%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Adverse Events/ Harms</td>
<td>2</td>
<td>2.4</td>
<td>3.0</td>
<td>0-3</td>
<td>36%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Bias/ Conflict of Interest</td>
<td>3</td>
<td>2.1</td>
<td>2.0</td>
<td>0-3</td>
<td>34%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Interpretation (Consistent with Data)</td>
<td>2</td>
<td>2.5</td>
<td>3.0</td>
<td>1-3</td>
<td>38%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Overall Evidence of Study</td>
<td>2</td>
<td>2.3</td>
<td>2.0</td>
<td>0-3</td>
<td>49%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Total % Score</strong></td>
<td><strong>73%</strong></td>
<td><strong>77.3%</strong></td>
<td><strong>39%</strong></td>
<td><strong>%</strong></td>
<td><strong>%</strong></td>
<td><strong>%</strong></td>
<td></td>
</tr>
</tbody>
</table>

*Note. (n= 53) - Fields are highlighted (and the corresponding category) if there is a ≥0.7 point difference between the expert and student scores, if the % if students assigning the same score as the experts falls below 60%, and if there were any incorrectly assigned N/As.*
MLE Rubric Matching Score Analysis. Tables 12-14 provide various data summaries of the student’s pre-lab cohort and case control MLE Rubric scores and post-lab RCT MLE Rubric score compared to the expert scores for each study type. These analyses attempt to present data using meaningful cut-off points, like study elements with greater than 60% matching student and expert scores (Table 12) to indicate student strengths and rubric elements with fewer than 50% matching student and expert scores to indicate student weaknesses (Table 13). Similar analyses were conducted using point differences between the expert score and the median and average student scores. Student scores were considered different from the experts if there was greater than or equal to 0.7 of a point difference (Tables 9-11) between the student average score and the expert consensus score and a greater than or equal to 1 point difference between the median student score and the expert score. The number 0.7, rather than 1, was selected for the average to accommodate the attenuation effects of averages and capture scores that trended towards a 1+ difference (0.7, 0.8 and 0.9 are closer to 1). These different analyses were conducted in an attempt to mitigate potential errors associated with analysis choices (median versus average) and assumptions (the matching score approach assumes the expert score is “correct” and that there is not more than one correct response) and identify trends across analyses. The results of each type of analysis are presented next.

Scores that differed from the expert score by greater than or equal to 0.7 of a point for each MLE Rubric section were highlighted and are distributed as follows: 48% (12/25) of the rubric sections in the case control study had greater than or equal to 0.7 point difference (median or average), 32% (9/28) of median or average sections for the cohort study had a difference and 41% (13/32) of the sections for the RCT had a 0.7 of a
point difference. This score difference serves as one potential measure of identifying where the third year Doctor of Pharmacy students are relative to the experts.

Another potential measure of identifying where the third year doctor of pharmacy students are relative to the experts is by identifying the percentage of students assigning the same score as the experts. Students had more matching scores for the cohort study compared to the case control and RCT, with 60% or more of the students matching in 10 of the cohort MLE Rubric fields compared to just 3 fields in the case control and 4 fields in the RCT (Table 12). Table 14 shows study elements with less than 50% match and indicates whether the student score was higher or lower than the expert score. The specific elements of the rubric and direction of the scoring plays an important role in understanding the errors in students’ evaluations of the study.
<table>
<thead>
<tr>
<th>Study Element</th>
<th>Cohort</th>
<th>Case-control</th>
<th>RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background(^1)</td>
<td>64%</td>
<td>---</td>
<td>---*</td>
</tr>
<tr>
<td>Objectives(^1)</td>
<td>77%</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Control/Comparison Group(^1)</td>
<td>64%</td>
<td>---</td>
<td>68%</td>
</tr>
<tr>
<td>Randomization 1 (From a Large Population)(^3)</td>
<td></td>
<td></td>
<td>62%</td>
</tr>
<tr>
<td>Measure of Exposure Assessment(^2)</td>
<td>---</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>Changes in Study Design(^1)</td>
<td>66%</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Outcomes Described(^1)</td>
<td>---</td>
<td>---</td>
<td>64%</td>
</tr>
<tr>
<td>Measures/Methods Describes/Justified(^1)</td>
<td>66%</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Sample Size (Power Calculations)</td>
<td>63%</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Sample Size (Enrolled and Completed)(^1)</td>
<td>77%</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Confounding Factors</td>
<td>61%</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Interim Analysis(^1)</td>
<td>86%</td>
<td>79%</td>
<td>---</td>
</tr>
<tr>
<td>Participant Flow(^1)</td>
<td>---</td>
<td>61%</td>
<td>---</td>
</tr>
<tr>
<td>Outcomes and Estimation(^1)</td>
<td>88%</td>
<td>---</td>
<td>70%</td>
</tr>
</tbody>
</table>

\(^1\) Element is applicable to all three studies
\(^2\) Element is applicable to all three studies
\(^3\) Element is applicable to the cohort and case-control study types
* --- indicates fewer than 60% of the students matched the expert score
<table>
<thead>
<tr>
<th>Study Element</th>
<th>Cohort</th>
<th>Case-control</th>
<th>RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background(^1)</td>
<td>---*</td>
<td>---</td>
<td>11% ↑(^1)</td>
</tr>
<tr>
<td>Objectives(^1)</td>
<td>---</td>
<td>9% ↑↑</td>
<td>8% ↑</td>
</tr>
<tr>
<td>Study Design(^1)</td>
<td>30% ↓</td>
<td>30% ↓</td>
<td>---</td>
</tr>
<tr>
<td>Generalizability(^1)</td>
<td>38% ↓</td>
<td>41% ↑</td>
<td>49% (same) (^2)</td>
</tr>
<tr>
<td>Participation Rate(^2)</td>
<td></td>
<td></td>
<td>38% ↓</td>
</tr>
<tr>
<td>Treatment Plan/ Intervention 1(^3)</td>
<td>38% (same)</td>
<td></td>
<td>32% ↑</td>
</tr>
<tr>
<td>Treatment Plan/ Intervention 2(^4)</td>
<td>32% ↑</td>
<td></td>
<td>30% ↑</td>
</tr>
<tr>
<td>Temporality(^5)</td>
<td>32% ↓</td>
<td></td>
<td>21%</td>
</tr>
<tr>
<td>Changes in Study Design</td>
<td>---</td>
<td>---</td>
<td>23% ↑</td>
</tr>
<tr>
<td>Measures/Methods Describes/Justified(^1)</td>
<td>---</td>
<td>23% ↑</td>
<td>47% (same)</td>
</tr>
<tr>
<td>Quality Enhancement(^1)</td>
<td>29% (same)</td>
<td>41% (same)</td>
<td>23% ↓</td>
</tr>
<tr>
<td>Sample Size (Power Calculations)(^1)</td>
<td>---</td>
<td>---</td>
<td>11% ↑↑(^4)</td>
</tr>
<tr>
<td>Sample Size (Enrolled and Completed)(^1)</td>
<td>---</td>
<td>20% ↑</td>
<td>30% ↑</td>
</tr>
<tr>
<td>Statistical Methods(^1)</td>
<td>45% ↓</td>
<td>39% (same)</td>
<td>---</td>
</tr>
<tr>
<td>Confounding Factors(^1)</td>
<td>---</td>
<td>27% ↑</td>
<td>21% ↑</td>
</tr>
<tr>
<td>Interim Analysis(^1)</td>
<td>---</td>
<td>---</td>
<td>8% ↑↑</td>
</tr>
<tr>
<td>Participant Flow(^1)</td>
<td>---</td>
<td>---</td>
<td>34% ↑</td>
</tr>
<tr>
<td>Key Data(^1)</td>
<td>45% ↑</td>
<td>39% (same)</td>
<td>43% ↓</td>
</tr>
<tr>
<td>Numbers Analyzed(^1)</td>
<td>---</td>
<td>43% ↓</td>
<td>---</td>
</tr>
<tr>
<td>Outcomes and Estimation(^1)</td>
<td>---</td>
<td>25% ↑</td>
<td>---</td>
</tr>
<tr>
<td>Ancillary Analyses(^1)</td>
<td>20% ↓</td>
<td>13% ↓↓</td>
<td>45% (same)</td>
</tr>
<tr>
<td>Adverse Events/ Harms(^2)</td>
<td></td>
<td></td>
<td>36% ↑</td>
</tr>
<tr>
<td>Bias/Conflict of Interest(^1)</td>
<td>---</td>
<td>---</td>
<td>34% ↓</td>
</tr>
<tr>
<td>Interpretation Consistent With Data(^1)</td>
<td>---</td>
<td>45% (same)</td>
<td>38% ↑</td>
</tr>
</tbody>
</table>
Table 13, continued

*--- indicates 40% or more of the students matched the expert score

1 Element is applicable to all three studies
2 Element is applicable to the RCT only
3 Element is applicable to the cohort and RCT
4↑↑ indicates that median student score was higher than the expert score, ↑↑ indicates greater than a 1.5 point difference
5(Same) indicates that the median score is the same as the experts’ consensus score

For the cohort study the four highest percentages of students assigning the same score as the experts in a single category of the MLE Rubric were 88%, 86%, 77% and 77%. These results occurred in the following MLE Rubric sections, respectively: the outcomes and estimation sub-section of the Results and Conclusions; the interim analysis sub-section of the Statistics section; the Introduction, objectives sub-section; and the Study Design/Methods, sample size enrolled and completed sub-section.

For the case control study, the two highest percentages of students assigning the same score in a single rubric category were 79% for the interim analysis sub-section of the Statistics section and 61% for the participant flow sub-section of the Results and Conclusions (both of which are N/A options). The next highest score match for the case control was 60% for the measure of exposure assessment. For the RCT the three highest percentages of students assigning the same score in a single rubric category were 70% for the outcomes and estimation category, 68% for the control/comparison group, and 64% for outcomes described. There are no MLE Rubric elements where greater than 60% of the students matched the expert for all three studies. Three sections of the MLE Rubric had a greater than or equal to 60% match for two of the three studies (highlighted in Table 12). These sections were control/comparison group, interim analysis, and outcomes and estimation.
Students seemed to do better in evaluating the cohort study compared to the case-control and RCT. More than 60% or more of the students matched in 10 of the MLE Rubric fields in the cohort study compared to just 3 fields in the case control and 4 fields in the RCT. Although these data from Table 12 are intended to show strengths, some instructors might not be satisfied if only 60% of the students demonstrated understanding of the material. Thus, a higher percentage cut-off point may be a better indicator of student strength. For a more robust analysis, Table 14 summarizes how this measure compares to the percentage of mean and median student scores that differed by greater than or equal to 0.7 points. For the most part these different analyses (average versus median versus matching scores) yielded similar results showing that students had the most difficulty evaluating a complex RCT. However, there were differences between the results of the analytic approaches used. For example, with the RCT data, if the cut off is greater than or equal to a 0.7 point difference between the students’ average scores and the expert scores, 34% of the MLE Rubric sections differ or may be weak. Using the less than 50% matching score criteria (where 50% or less of the students matched the expert score), then 63% of the MLE Rubric sections differ or may be weak. This indicates a difference between the approaches.

A natural question is which approach should be used? There are potential problems with any of the approaches. For example, the matching approach assumes that the expert answer is “correct” and that there is only one right answer. In some cases this may be true and in others an MLE Rubric score of a two may be just as correct as a three. However, the matching approach is much more fine grained because averages and medians lose information and thus, may not detect areas of student misunderstanding. For
example, many students may score a section of the article with a zero or a three, which would place the median around a 1.5 or a two. Both a zero or three could be clearly wrong responses. If the experts scored that section with a one, then the difference would not be picked up with the median or average approach.

Table 14 also shows areas where using the median as an indicator is weak. There are ten examples where the median is the “same” as the expert score, but only a small percentage of the students actually matched the expert score, with 21% to 49% matching.

Table 14

<p>| Percentage of Rubric Sections Differing From or Matching Expert Scores |
|---------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>% of Rubric Sections with ≥ 0.7 Difference in Score Compared to the Expert Score (Average)</th>
<th>% of Rubric Sections with ≥ 0.7 Difference in Score Compared to the Expert Score (Median)</th>
<th>Total % of Rubric Sections with ≥ 0.7 Difference in Score Compared to the Expert Score (Median &amp; Average)</th>
<th>% of Rubric Sections where &lt;50% of the Students’ Scores that Match the Expert Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case Control</td>
<td>36% (9/25)</td>
<td>32% (8/25)</td>
<td>48% (12/25)</td>
<td>52% (13/25)</td>
</tr>
<tr>
<td>Cohort</td>
<td>29% (8/28)</td>
<td>25% (7/28)</td>
<td>32% (9/28)</td>
<td>32% (9/28)</td>
</tr>
<tr>
<td>RCT</td>
<td>34% (11/32)</td>
<td>41% (13/32)</td>
<td>41% (13/32)</td>
<td>63% (20/32)</td>
</tr>
</tbody>
</table>

1Excludes the N/A fields

Another area that is indicative of the student’s ability to evaluate the medical literature is the number of N/As that were incorrectly assigned. There are three sections in the rubric where N/A may be applicable (blinding, interim analyses, and participant flow). For the cohort study, two of the three possible N/A categories were assigned an
N/A by the experts (blinding and interim analyses). All three of the N/A categories were assigned an N/A by the experts for the case control study and none of the categories were assigned an N/A for the RCT.

For both the cohort and case control studies, students had the most difficulty identifying that the blinding should have been given an N/A. In both cases, 23 students (41%) assigned the study a score instead of assigning an N/A. Students also had difficulty identifying whether the participant flow should have been rated N/A. In the cohort study, participant flow should have been given a score and 11 (20%) students assigned it an N/A. Conversely, in the case control study, students should have assigned an N/A to participant flow, but 23 (41%) assigned a score.

For the cohort study, in 28% of the fields (7 out of 25 possible fields), students incorrectly assigned/did not assign an N/A. Four of those fields, eligibility criteria, treatment plan/intervention 1, treatment plan/intervention 2, and changes in study design, did not have N/A as an option but still 2 to 4 students placed an N/A. For the case control study, 18% (5 out of 25) of the fields were incorrectly labeled for N/A. Two of those five fields, changes in study design and sample size, did not have N/A as an option, but four and one students respectively, still placed an N/A in the field.

Higher numbers of students had difficulty with blinding. Although 59% of the students matched the expert score (which was N/A) for blinding, 41% (23) students did not identify that the section needed to be labeled N/A and they instead provided a score.

**Article Ranking.** After the students evaluated the articles using the MLE Rubric, they were asked to rank them by assigning the number 1 to the best study/article, 2 for the second best and 3 for the lowest quality study. Students were asked to rank the articles
twice, once prior to the lab considering just the cohort and case control and then once again at the end of the activity ranking all three articles relative to each other. In the first ranking, 76% (53) identified the cohort as a better study than the case control. At the end of the activity, when the students considered all three articles, 42% ranked the cohort as number one, 42% ranked the RCT as number one, and 15% ranked the case control study as number one. The average student MLE Rubric scores, which were 77% for the RCT, 75% for the cohort and 71% for the case control. Although the difference in the average MLE Rubric scores is comparatively small, if based solely on the average total study score, the rank order would be RCT, cohort, and case control. Table 15 shows article ranking prior to lab and following. It should be noted that the ranking done at the end of lab was completed prior to the instructor revealing the expert scores. The expert scores showed a distinct spread/difference between the quality of the studies (87% cohort, 73% RCT and 65% case control).

Table 15

Pre- and Post-Activity Article Ranking

<table>
<thead>
<tr>
<th>Rank</th>
<th>Prior to Lab (Having Completed Rubric)</th>
<th>End of Lab (Prior to Instructor Revealing Expert Ranks and Scores)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case Control</td>
<td>Cohort</td>
</tr>
<tr>
<td>1</td>
<td>24%</td>
<td>76%</td>
</tr>
<tr>
<td>2</td>
<td>76%</td>
<td>24%</td>
</tr>
<tr>
<td>3</td>
<td><strong>59%</strong></td>
<td><strong>8%</strong></td>
</tr>
</tbody>
</table>

Student Average Total Study Score | 71% | 75% | 77%

Expert Consensus Score | 65% | 87% | 73%
**Crux Vocabulary.** Crux vocabulary refers to words or concepts that must be understood in-depth to be able to evaluate the quality of the study. Students were instructed to select four terms that are essential to understanding and evaluating the study and to provide a brief, specific description about how understanding that term or concept can impact the ability to evaluate the study.

Fifty-six students submitted four crux vocabulary words each for the cohort and case control studies, for a total of 224 submissions, 209 of which were analyzed because not all were complete (i.e. they were left blank) and several students copied each other’s work and provided the same responses for their crux words. In those cases only one of the two was used in the analysis. Of the 209 submissions there were 57 distinct crux vocabulary words submitted by the students (Appendix F). The words that received the highest number of submissions are listed below. If they match with a crux identified by the experts, it is indicated with an asterisk (*).

- Charlson Co-Morbidity Index (22 students)
- Cox models /Cox Proportional Hazard Model (18 students)
- cardiovascular death/cardiovascular event/*MACE/coding for
  *MI/reinfarction (12 students)
- *ICD9 Codes (13 students)
- CYP2C19/Cytochrome P450 (9 students)

The remainder of the crux words were submitted by one to five students. The experts identified 11 different crux vocabulary words: adjudication; adjusted odds ratio; a-priori power analysis; aspirin dose; bleeding event; cardiovascular events; current use; ICD-9/10 codes; myocardial infarction (MI); MACE; and percutaneous coronary
intervention (PCI). Fifty-five percent of the expert identified crux words were also identified by the students (adjusted odds ratio, current use, ICD-9 codes, myocardial infarction, MACE, and percutaneous coronary intervention) (Table 16). Many of the other words identified by the students are important words but they are not “crux” words. Table 17 also shows the distribution of students in the various qualitative categories for each matched crux vocabulary.

Table 16

**Matched Crux Vocabulary: Students and Expert Selections**

<table>
<thead>
<tr>
<th>Matched Crux Vocabulary</th>
<th>Number of Students Identifying the Word</th>
<th>Category of Student Descriptive Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjusted odds ratio</td>
<td>3</td>
<td>1 definition, 1 partial</td>
</tr>
<tr>
<td>current use</td>
<td>2</td>
<td>1 weak, 1 partial</td>
</tr>
<tr>
<td>ICD-9 codes</td>
<td>13</td>
<td>4 weak, 2 definition, 6 partial, 1 reason</td>
</tr>
<tr>
<td>myocardial infarction</td>
<td>1</td>
<td>definition</td>
</tr>
<tr>
<td>MACE</td>
<td>8</td>
<td>2 weak, 3 definition, 3 partial, 1 reason</td>
</tr>
<tr>
<td>percutaneous coronary intervention</td>
<td>5</td>
<td>1 weak, 3 partial, 1 reason</td>
</tr>
</tbody>
</table>

In analyzing definitions/explanations, responses provided in students were sorted into the following categories – weak, definition, partial and reason. The categories are defined and listed below from the lowest to highest response quality. If a response could fall into more than one category, it was placed in the “highest” category. For example, if a student provided both a definition and partially addressed “how understanding that term or concept can impact the ability to evaluate the study,” it was categorized as “partial.” Examples from each of the categories from one of the matched crux vocabulary are provided.
• **Weak** = Student provided a weak response that had very little substance or meaning. S/he did not provide a definition nor adequately described “how understanding that term or concept can impact the ability to evaluate the study”
  
  o EXAMPLE: ICD9 Codes - One can get confused if they [sic] do not know what this is, or which procedure a certain code pertains to.

• **Definition** = Student provided a definition of the word, but did not address the question “how understanding that term or concept can impact the ability to evaluate the study.”
  
  o EXAMPLE: ICD9 Codes - The ICD-9-CM is a classification system, maintained by the National Center for Health Statistics and the Center for Medicare and Medicaid, that can be used for billing and other medical purposes. The study used ICD-9-CM codes to track reinfarction incidence in each cohort.

• **Partial** = Statement partially addressed “how understanding that term or concept can impact the ability to evaluate the study,” but the response was incomplete or slightly missed the point.
  
  o EXAMPLE: ICD9 Codes - Researches that pulled out the data from the system have to know the numbers and information correlated with each variable of the study so they could pull it out of the system for the study. The information was projected to us in a chart and if one did not know what they were looking at then they would look at the chart and see numbers that meant nothing and this would possible raise the question of where the researches got the information they were presenting.
- **Reason** = Student adequately described “how understanding that term or concept can impact the ability to evaluate the study.”
  
  - EXAMPLE: ICD9 Codes - These codes are used to provide a diagnosis and describe a clinical event. These codes are often complex and seem to be very inclusive in the cohort study. This may lead to confounding.

  Things can be coded wrong.

  The distribution of crux vocabulary across the categories is relatively equal: 23% were weak; 24% provided only definitions; 34% of the responses partially addressed the question “how understanding that term or concept can impact the ability to evaluate the study”; and 18% provided a “reason.”

  A correlation analysis was performed in an attempt to discern whether students’ ability to identify and/or describe a crux vocabulary term has any connection with MLE Rubric scores that more closely resemble expert scores. The following hypotheses were tested using IBM SPSS Statistics Version 20 running a two-tailed non-parametric bivariate correlation test with Kendall’s Tau and Spearman’s Rho correlation coefficients:

  - students with more crux vocabulary words that matched with the expert’s crux words would have more matched MLE rubric section scores that matched with the experts; and
  
  - students with higher-order crux vocabulary responses (at the partial and reasoning level) would have more matched MLE rubric section scores that matched with the experts;

  Both types of correlation coefficients were consistent with each other in all tests.
The tests paired the number of matching crux vocabulary with the percent of student rubric sections matched with the experts for each of the three study types (case control, cohort and RCT) and the number of responses in each crux vocabulary rating category (weak, definition, partial, and reason) with the percent of student rubric sections matched with the experts for each of the three study types.

None of the comparisons showed any statistically significant correlations, with the exception of one test comparing the crux vocabulary responses in the “partial” category and the percent matched MLE rubric sections for the cohort study (Table 17). The test showed a weak negative correlation (-0.238 Kendall’s Tau correlation coefficient, p= 0.03; -0.289 Spearman’s rho correlation coefficient, p=0.036). The nature of the potential inverse relationship can be seen the scatter plot (Figure 6).

In order to further explore the question about potential relationships between crux vocabulary and ability to evaluate the medical literature, the number of words in the higher order categories, “partial” and “reason,” were combined to increase the number of responses for comparison. This single category was paired with the percent of student rubric sections matched with the experts for each of the three study types and no correlations were found (Table 17).
Table 17  
*Pairwise Bivariate Correlation Tests for Crux Vocabulary and the Percentage of Categories*

<table>
<thead>
<tr>
<th>Combinations (Crux Vocabulary Category + % Matched MLE Rubric Sections Study Type)</th>
<th>Kendall’s Tau Correlation Coefficient</th>
<th>Significance (Two-tailed)</th>
<th>Spearman’s Rho Correlation Coefficient</th>
<th>Significance (Two-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak + Case Control</td>
<td>-0.167</td>
<td>0.138</td>
<td>-0.219</td>
<td>0.115</td>
</tr>
<tr>
<td>Weak + Cohort</td>
<td>0.105</td>
<td>0.348</td>
<td>0.127</td>
<td>0.363</td>
</tr>
<tr>
<td>Weak + RCT</td>
<td>0.065</td>
<td>0.568</td>
<td>0.077</td>
<td>0.594</td>
</tr>
<tr>
<td>Definition + Case Control</td>
<td>0.075</td>
<td>0.504</td>
<td>0.095</td>
<td>0.500</td>
</tr>
<tr>
<td>Definition + Correlation Coefficient</td>
<td>0.064</td>
<td>0.562</td>
<td>0.083</td>
<td>0.555</td>
</tr>
<tr>
<td>Definition + RCT</td>
<td>0.087</td>
<td>0.441</td>
<td>0.107</td>
<td>0.455</td>
</tr>
<tr>
<td>Partial + Case Control</td>
<td>-0.143</td>
<td>0.197</td>
<td>-0.180</td>
<td>0.198</td>
</tr>
<tr>
<td>Partial + Cohort</td>
<td>-0.238</td>
<td>0.030*</td>
<td>-0.289</td>
<td>0.036*</td>
</tr>
<tr>
<td>Partial + RCT</td>
<td>-0.117</td>
<td>0.301</td>
<td>-0.136</td>
<td>0.347</td>
</tr>
<tr>
<td>Reason + Case Control</td>
<td>0.109</td>
<td>0.345</td>
<td>0.138</td>
<td>0.335</td>
</tr>
<tr>
<td>Reason + Cohort</td>
<td>0.014</td>
<td>0.903</td>
<td>0.015</td>
<td>0.915</td>
</tr>
<tr>
<td>Reason + RCT</td>
<td>0.002</td>
<td>0.984</td>
<td>0.003</td>
<td>0.986</td>
</tr>
<tr>
<td>Partial and Reason Combined + Case Control</td>
<td>-0.016</td>
<td>0.886</td>
<td>-0.015</td>
<td>0.917</td>
</tr>
<tr>
<td>Partial and Reason Combined + Cohort</td>
<td>-0.182</td>
<td>0.097</td>
<td>-0.224</td>
<td>0.114</td>
</tr>
<tr>
<td>Partial and Reason Combined + RCT</td>
<td>-0.059</td>
<td>0.599</td>
<td>-0.072</td>
<td>0.622</td>
</tr>
<tr>
<td>No. of Crux Vocabulary Matching Experts + Case Control</td>
<td>-0.123</td>
<td>0.278</td>
<td>-0.149</td>
<td>0.283</td>
</tr>
<tr>
<td>No. of Crux Vocabulary Matching Experts + Cohort</td>
<td>-0.107</td>
<td>0.342</td>
<td>-0.138</td>
<td>0.319</td>
</tr>
<tr>
<td>No. of Crux Vocabulary Matching Experts + RCT</td>
<td>-0.121</td>
<td>0.297</td>
<td>-0.148</td>
<td>0.300</td>
</tr>
</tbody>
</table>

*Statistically significant at 0.05 level*
Clinical Conclusion. Students were asked to provide a clinical conclusion at two different points in the activity. The first was based upon the cohort and case control articles alone and the second considered all three articles (students had been exposed to the expert’s analysis of the cohort and case control, but NOT the expert’s analysis of the RCT and NOT the experts’ clinical conclusion). The experts’ clinical conclusion was:

Based on the three articles there appears to be a potential for an interaction between PPIs and clopidogrel which can increase patients’ risk for thrombotic complications. If the patient can tolerate another medication class such as histamine blockers, a change should be considered. If a patient was at sufficient risk of GI bleeding (previous GI bleed, advanced age, concomitant anticoagulation therapy, corticosteroid use) and required use of dual antiplatelet therapy (plavix and asa), then I would recommend the use of a PPI and would not give omeprazole and clopidogrel together (other PPIs would be OK). That decision is only using the three pieces of literature we evaluated, if considering the larger body of evidence, there are more inconsistencies in the results and the association between these two drug classes, in terms of an interaction that increases cardiovascular risk, appears to be smaller.
There were a total of 55 responses in the first clinical conclusion and 56 in the final clinical conclusion. The decisions were sorted into 8 categories plus one subcategory.

The clinical conclusion using only the cohort and case control studies had 51% of the responses falling into 2 or more categories and 7% of the responses did not provide a clinical conclusion (Table 18). The categories are:

- More data (MD)
- Do not use/avoid clopidogrel and PPI together (NoCPPI)
- Not sure (NS)
- Other approach (OA)
- Use an H\(_2\) antagonist (H\(_2\)A)
- Use pantoprazole (P)
- Use any PPI/omeprazole (PPI)

For the first clinical conclusion based upon the cohort and case control alone, the two ends of the response spectrum were 22% of the students said they would still use any PPI/omeprazole without any mention of other alternatives, while 16% simply said they would not use PPIs, but they did not offer an alternative (Table 18). The majority of the initial clinical conclusions (47%) were to use an H\(_2\) antagonist, either to entirely replace PPIs (16%) or as a preference to PPIs, but they would still consider the use of PPIs (18%). Thirty-one percent said they would use pantoprazole as an alternative choice either after first considering an H\(_2\) antagonist or as a preference to other PPIs.
After students read the RCT and had all three articles on which to base their clinical conclusion (Table 19), the two ends of the response spectrum changed, although most of the shifts were not statistically significant (one was). Following the RCT 36% of the students said they would still use any PPI/omeprazole without any mention of other alternatives (it was 22% prior to reading the RCT, not significant), 32% said they would use an H₂ antagonist (which was previously 47%, not significant), and 24% explicitly said they would not use PPIs at all (previously it was 16%, not significant). The majority of the final clinical conclusion (36%) was to use omeprazole with clopidegrel, which is more consistent with the RCT than the other two studies. One change in the clinical conclusion that was significant was a shift away from a decision that listed multiple options or a combination treatment approach to a singular option approach (43% to 23%, p=0.0232). The multiple option approach was more consistent with the expert clinical conclusion.

Table 18

Initial Clinical Conclusion using the Cohort and Case Control Studies Only

<table>
<thead>
<tr>
<th>Not Sure (NS)</th>
<th>Other Approach (OA)</th>
<th>More Data (MD)</th>
<th>No PPI (noPPI)</th>
<th>Use an H₂ Antagonist (H₂A)</th>
<th>Use Pantoprazole (P)</th>
<th>Use PPI in general/omeprazole (PPI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7%</td>
<td>4%</td>
<td>5%</td>
<td>16%</td>
<td>47%</td>
<td>31%</td>
<td>22%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16% listed these options</td>
<td></td>
<td>9% listed these options</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18% listed these options</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Note n=55, students could select more than one response)
Table 19

*Final Clinical Conclusion using all Three Studies*

<table>
<thead>
<tr>
<th>Not Sure (NS)</th>
<th>Other Approach (OA)</th>
<th>More Data (MD)</th>
<th>No PPI (noPPI)</th>
<th>Use an H₂ Antagonist (H₂A)</th>
<th>Use Pantoprazole (P)</th>
<th>Use PPI in general/ omeprazole (PPI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0%</td>
<td>3%</td>
<td>24% (p=0.2937)</td>
<td>32% (p=0.1051)</td>
<td>32%</td>
<td>36% (p=0.1029)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Note n=56, students could select more than one response, there is no statistically significant change at a 95% confidence level, 2-tailed $z$-test of sample proportions)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In addition to collecting data on students’ clinical conclusions, they were also asked to rate their level of confidence in their initial and final clinical conclusions (Tables 20 and 21). For the initial decision, there was a nearly even split between somewhat confident (44%) and confident (42%). In the final clinical conclusion the percentage of matched students marking “confident” increased to 62% (this is a statistically significant change, $p=0.033$). This shows that students may have difficulty discerning between a strong or weak clinical conclusion and the potential influence of bias towards RCTs as a result of the hierarchy of evidence.
Table 20

Confidence Ratings for Initial and Final Clinical Conclusions

<table>
<thead>
<tr>
<th></th>
<th>Initial Clinical Conclusion (n=55)</th>
<th>Final Clinical Conclusion, Matched Responses Initial and Final (n=47)</th>
<th>Final Clinical Conclusion, All Responses (n=67)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited Confidence</td>
<td>4% (2)</td>
<td>0%</td>
<td>3% (2)</td>
</tr>
<tr>
<td>Somewhat Confident</td>
<td>44% (24)</td>
<td>28%</td>
<td>27% (18)</td>
</tr>
<tr>
<td>Confident</td>
<td>41% (23)</td>
<td>62%*</td>
<td>55% (37)</td>
</tr>
<tr>
<td>Very Confident</td>
<td>6% (11%)</td>
<td>6%</td>
<td>15% (10)</td>
</tr>
</tbody>
</table>

*statistically significant, p=0.033

Table 21

Change in Decision Confidence Between Initial and Final Clinical Conclusions

<table>
<thead>
<tr>
<th>Change in Confidence</th>
<th>Percent (n=47 matched pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase</td>
<td>28%</td>
</tr>
<tr>
<td>Decrease</td>
<td>13%</td>
</tr>
<tr>
<td>No Change</td>
<td>60%</td>
</tr>
</tbody>
</table>

Scaffold Contributions Survey

The final set of data collected was student self-perceptions about the contributions of the scaffolds to their learning (Table 22). Students completed the Scaffold Contributions Survey at the end of the activity (Appendix G). In the first question students were asked to rate the following instructional elements (the MLE Rubric by itself, small group peer discussion about the RCT, the large class discussion around the MLE rubric and articles, the crux vocabulary assignment and the class discussion about...
the crux vocabulary) on how well they supported learning around evaluating different
types of medical literature and they elected one of four responses (no contribution to my
learning, a little contribution to my learning, moderate contribution to my learning, high
contribution to my learning).

The large class discussion was perceived as the scaffold that contributed the most
to their learning with 56% of the students selecting “high contribution to my learning.”
At 38%, the MLE Rubric received the second highest number of students selecting the
“high contribution to my learning” and the small group discussion came in third with
33% selecting the highest learning option. All three of these scaffolds were distributed at
the higher end of the scale with the majority of the students rating them as “high
contribution to my learning” or moderate contribution to my learning.

All scaffolds had at least two students select the “no contribution to my learning
option” with the crux vocabulary receiving the highest percentage of students selecting
this option (11%). The crux vocabulary assignment and the class discussion about the
crux vocabulary were primarily distributed between “a little” (46% and 34%) and
“moderate contribution to my learning” (38% and 40%) respectively. The class
discussion had a comparatively higher percentage selecting “high contribution to my
learning with 23% compared to 5%.
Table 22

*Scaffold Contributions Survey Results*

<table>
<thead>
<tr>
<th>Rate the following instructional elements on how well they supported your learning around evaluating different types of medical literature.</th>
<th>No contribution to my learning</th>
<th>A little contribution to my learning</th>
<th>Moderate contribution to my learning</th>
<th>High contribution to my learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>The MLE Rubric by itself</td>
<td>5% (3)</td>
<td>19% (12)</td>
<td>39% (25)</td>
<td>38% (24)</td>
</tr>
<tr>
<td>The small group (3-4 people) peer discussion about the RCT</td>
<td>8% (5)</td>
<td>16% (10)</td>
<td>43% (27)</td>
<td>33% (21)</td>
</tr>
<tr>
<td>The large class discussion around the MLE rubric and articles</td>
<td>6% (4)</td>
<td>5% (3)</td>
<td>33% (21)</td>
<td>56% (35)</td>
</tr>
<tr>
<td>The Crux Vocabulary assignment</td>
<td>11% (7)</td>
<td>46% (29)</td>
<td>38% (24)</td>
<td>5% (3)</td>
</tr>
<tr>
<td>The class discussion about the crux vocabulary</td>
<td>3% (2)</td>
<td>34% (21)</td>
<td>40% (25)</td>
<td>23% (14)</td>
</tr>
</tbody>
</table>

(n= 62 or 64)

Since approximately 2/3 of the assignment was completed outside of class, students were asked to identify whether they worked alone for the outside, pre-lab activities or with others (Table 23). A large majority (83%) of the respondents indicated they worked alone, while 15% worked both alone and with student peers. Only one student completed all of the pre-lab work with peers. This information limits a potential major confounder in the study (peer influence on the work) and increases the confidence that the majority of the data indicate that the student work product is primarily a
reflection of two things: (1) individual background knowledge and (2) the contribution of the MLE Rubric and crux vocabulary scaffolds to learning.

Fourteen of the respondents for this question provided comments elaborating on working alone, commenting about the length of the activity, or making other suggestions not related to working alone/with others. The comments that related to working alone were categorized as follows:

- O+ = working with others helps me/would have been better to work with others;
- A+ = prefer working alone;
- AD = difficult working alone.

Twenty-nine percent of the comments indicated a preference to working alone, at least initially. The majority of those also stated that after they worked alone, working with others afterwards helped them in some way. For example “It was helpful to work by myself and follow it up working with 2-3 peers before class.” Fifty percent of the comments were rated as O+ (again two of which also stated a preference for working alone initially), indicating a preference to work with others. Examples of comments included: “I worked alone and I always think working with other students helps me, it would have especially helped me understand the CRUX word;” “I worked alone, however, once i (sic) got into lab and discussed the trials in a group setting, I felt I understood them much better;” and “A friend and I took turns reading the articles, that way we payed (sic) attention and did not get discouraged.” One comment was categorized differently, as it focused on it being difficult to work alone and did not explicitly belong in either of the other categories. That comments was, “Ranking and grading the tirals (sic) alone was somewhat overwhelming.”
Table 23

Student Comments Related to Working Alone

<table>
<thead>
<tr>
<th>For the pre-lab activities did you work by yourself or with other students? (Don’t worry there is no right or wrong answer to this – i.e. you won’t get in trouble – it just helps us to better understand how students work outside of class and the contributions of peer conversations plus the MLE Rubric or the MLE Rubric alone). Feel free to add comments on this topic if you would like.</th>
<th>Response Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>I worked alone on the pre-lab activities</td>
<td>83%</td>
</tr>
<tr>
<td>I worked alone and with student peers</td>
<td>15%</td>
</tr>
<tr>
<td>I did all of the pre-lab work together with student peers</td>
<td>2%</td>
</tr>
</tbody>
</table>

4. Comments Response Categories (n=14)

| O+ = working with others helps me/would have been better to work with others | 7 (5 of which are O+ only) |
| A+ = prefer working alone | 4 (2 of which are A+ only) |
| AD = difficult working alone. | 1 |
| Comments not related to working alone (general comments about the activity, which are addressed in the next question) | 4 |

(n=56, 13 comments)

Limitations of the Study

There were several limitations of the study. Although there were 89 students in the class, only 56 consented to the study. Thus, conclusions are being drawn on a limited sample. It is possible that the students that consented constituted a proportionally “weaker” or “stronger” group. Also, although students received points towards their grade for completing the assignment, which was intended to be a motivator to complete the work, it was credit/no credit as opposed to a grade. This may have resulted in some students putting in minimal effort into the work product. Another attempt to improve student effort in the activity was not holding the 2-hour whole class session (session with
89 students) during the week of the activity. Students only had to attend the 3-hour lab session (session with 28-29 students) where parts of the activity occurred. There is no guarantee that students utilized the extra time to carefully complete the work. All of these factors may increase the risk of erroneous conclusions about actual student ability. Other limitations of the study included small inconsistencies in implementation of the activity between class sections (in sessions two and three the instructor/expert invited more students to share their crux words) and, according to the post-activity survey, a few students indicated they worked together outside of class.
Chapter 5

Discussion

The goal of this study was to test a scaffolded instructional model in a complex Evidence Based Medicine (EBM) lesson where students evaluated three different medical studies (RCT, case-control, and cohort) on the same topic and used those studies to reach an evidence-based conclusion. The hypothesis was that the students’ literature evaluation skills and subsequent application of the literature to address a clinical question would more closely approximate the experts following implementation of the model. The results do not fully support the hypothesis. The conclusions were: (1) third-year doctor of pharmacy students at the college have a limited ability to evaluate medical literature of varying qualities and types and conflicting conclusions; (2) prior to reading the RCT, students’ initial clinical conclusions more closely resembled the experts’, potentially indicating an unbalanced influence of the RCT, either from RCT bias/preconceptions or a lack of skills transfer in evaluating the RCT; and (3) the instructional model needs further development by adding explicit instructional scaffolding around the Medical Literature Evaluation (MLE) Rubric, vocabulary, and directly addressing student preconceptions/biases.

EBM is the practice of utilizing current, high quality scientific evidence in individual patient care, be it in medicine, nursing, or pharmacy and is necessary for today’s health care practitioners to provide a high level of patient care. There are five core EBM components, each of which requires its own skills, training, and definitions regarding what constitute best practice. The EBM components are: (a) accurately assessing a patient; (b) asking appropriate clinical questions to implement a meaningful
literature search; (c) acquiring a body of high quality literature; (d) appraising the literature on validity, importance and usefulness (which is the focus of this dissertation); and (e) interpreting and applying the evidence to the clinical issue (also the focus of this dissertation) (retrieved from JAMAevidence.com).

The ability to appraise “the literature on validity, importance and usefulness” and draw clinical conclusions from multiple and different kinds and qualities of studies requires in-depth knowledge of the medical topic as well as research methodologies. Methodologies include things like sampling methods, sample numbers requirements, inclusion/exclusion criteria, blinding, and other ways systematic bias and/or error can be addressed or minimized through the study design. It also includes how data can (and should) be analyzed using appropriate statistical methods and reasonable interpretation of statistics for that study type. In addition to the need to understand the research methods, medical literature evaluation requires knowledge of the medical research topic (e.g., the mechanism of action of a specific drug in a specific disease state) and the vocabulary and procedures (such as common lab tests) associated with that medical issue.

Evaluating the medical literature and reaching an accurate and unbiased clinical conclusion based upon evidence is a complex and multi-layered process that is historically under-taught in the medical curricula (Green, 2000; Kuhn, et al., 2005). Over the past several years the COP has increased EBM in its curriculum and modified its approaches to include more active learning as opposed to primarily lecture-based approaches, to improve learning around EBM.

In the second year of pharmacy school at the COP, students take a required pharmacy informatics and research course in which they learn about the different types of
medical studies (e.g., meta-analyses/reviews, RCT, case-control, cohort, and case studies), some basic statistical review, and where they complete the first literature analysis activity for an RCT using the Clinical Literature Evaluation (CLE) Rubric (developed by me) (Dawn, et al. 2011). At the end of the semester, students complete a group literature review paper on a pharmacy topic.

In addition to the semester-long course work in pharmacy informatics and research, students participate in 1-2 journal club activities in their second year pharmacy lab (lab). In the third year, students are given a rigorous assignment to reach a clinical conclusion on a pharmacy topic from a list of topics using current literature.

Once students enter clinical rotations in their fourth year of pharmacy school they are expected to apply their skills evaluating medical literature and to reach an accurate and unbiased clinical conclusion, especially for current topics that might not be in the review literature.

The third-year lab instructor and numerous faculty who teach in fourth-year rotations anecdotally noticed (from the third year lab activity referred to above) that despite the improved second-year coursework, and additional journal club activities, students still appeared to have difficulty identifying high quality studies on current topics and that the clinical conclusions students reached were often inaccurate, incomplete and/or based upon low-quality or highly flawed studies.

After discussion with the third-year lab instructor, it was decided that students needed additional, targeted training in evaluating medical literature and reaching a clinical conclusion. As a result, I developed this multiple literature evaluation activity to build off of the second year RCT literature evaluation activity, which was previously
shown to be an effective instructional model and improved students’ abilities in evaluating an RCT (Dawn et al., 2011).

The goal of this study was to test a scaffolded instructional model in a complex Evidence Based Medicine (EBM) lesson where students evaluated three different medical studies (RCT, case-control, and cohort) on the same topic and used those studies to reach an evidence-based conclusion. The hypothesis was that the students’ literature evaluation skills and subsequent application of the literature to address a clinical question would more closely approximate the experts following implementation of the model.

The methods used for this research included two surveys (a characterization survey and scaffold contributions survey) and analysis of student work compared to experts before and after the implementation of certain scaffolds (student MLE Rubric scores and clinical conclusions).

There are three main conclusions as a result of this research:

- Conclusion One - Third-year doctor of pharmacy students at the COP have a limited ability to evaluate medical literature of varying qualities and types and conflicting conclusions;
- Conclusion Two - Prior to reading the RCT, students’ initial clinical conclusions more closely resembled the experts’, potentially indicating an unbalanced influence of the RCT, either from RCT bias or a lack of skills transfer in evaluating the RCT; and
- Conclusion Three – The instructional model needs further development by adding explicit instructional scaffolding around the Medical Literature
Evaluation (MLE) Rubric, vocabulary, and directly addressing student preconceptions/biases.

This chapter is organized around each conclusion and highlights the key data that support these conclusions. The chapter concludes with the implications for EBM more broadly and teaching EBM more specifically in the pharmacy curriculum and makes recommendations for future research.

**Conclusion One**

The first conclusion, which is third-year doctor of pharmacy students at the COP have a limited ability to evaluate medical literature of varying qualities and types and conflicting conclusions, is supported by multiple data points from the MLE Rubric and the crux vocabulary.

**Student challenges with analyzing the medical literature – the MLE Rubric.**

Because this activity had students read and evaluate three different types of medical literature on the same topic, with each study differing in quality and one of the studies having a different conclusion, it provided a comprehensive measure of a COP third-year student’s ability to assess the quality of particular studies and reach a clinical conclusion from conflicting data.

Even though:

- students claimed to have read many primary literature studies (53% claim to have read more than 21 studies);
- they took an informatics and research course in their second year and completed a similar, but much simpler, literature evaluation activity on a single RCT and have participated in journal clubs; and
they are in their third year of pharmacy school preparing to enter rotations where they will need to apply literature evaluations skills;

dthis research showed that the students are still developing their literature evaluation skills and their ability to reach an unbiased clinical conclusion using multiple, conflicting studies.

The MLE Rubric and comparison of student scores with expert scores provided a substantive amount of data in support of this conclusion (Chapter 4). For example, the average total percent score students assigned to the three studies differed from the expert consensus scores. The student scores also showed very little discrimination between study qualities. The average MLE Rubric scores students assigned to the studies were 71%, 75%, and 77% for the case-control, cohort and RCT, respectively. Compare this to the expert scores, which were 65% for the case-control, 87% for the cohort study, and 73% for the RCT.

The experts were not only able to differentiate between the quality of the studies, as seen by the score spread (65%-87%), they also scored the cohort study much higher than the default “gold standard” RCT (the experts assigned 87% for the cohort versus 73% for the RCT). Comparatively, the student MLE rubric score range was clustered around 75% (71%-77%) and highest score assigned by the students was for the RCT (77% for the RCT and 75% for the cohort).

Although the expert score of 73% for the RCT appears to be similar to the average student score for the RCT (77%), detailed analysis of the rubric subsection scores showed a very large discrepancy between student evaluation of the RCT and expert evaluation. Students performed the best matching the expert scores in the cohort study
where 36% of the MLE Rubric Fields had greater than 60% of the students matching the expert score. This is compared to just 12% fields in the case-control and 13% of the fields in the RCT.

**Student challenges with the crux vocabulary.** Students also had difficulty with crux vocabulary. Either they did not understand the concept and/or their skill level limited them from identifying true crux terms. The majority of the students’ “crux” terms did not match the expert’s crux terms. The experts identified eleven crux vocabulary terms, five of which were matched by very few students. The ICD-9 codes had the greatest number of matches with 13 of the 56 students (23%) listing that as one of their crux vocabulary. The remaining matches had between one and eight students (2-14%) per matching term. It appears that students were more likely to choose unfamiliar words, rather than words that played a critical role in evaluating the study.

When analyzing the quality of the crux vocabulary descriptions provided by the students, the majority of the descriptions did not meet the basic expectations for the “reason” level higher-order thinking category. The sorting categories were “weak,” “definition,” “partial,” and “reason.” The weakest/lower-order categories are “weak” and “definition,” where students gave little thought to the term description or simply copied and pasted a definition.

The “partial” description category showed some direction toward a higher-order response, and the responses that were categorized under “reason” met at least the basic expectations of the crux vocabulary description. Forty-seven percent of the descriptions the students provided for the matched expert terms fell into the “weak” or “definition”
categories and 34% of the responses fell under the “partial” category. Only 18% of the descriptions were categorized into “reason.”

One measure of the crux concept hypothesized that if students could correctly identify crux vocabulary and describe it at a higher level they would be able to evaluate a study more like an expert. Correlation analyses testing the following hypotheses were conducted:

- students with more crux vocabulary words that matched with the expert’s crux words would have more MLE rubric section scores that matched with the experts; and
- students with higher-order crux vocabulary responses (at the partial and reasoning level) would have more MLE rubric section scores that matched with the experts;

The analysis found no significant results with the exception of one test, and that showed a very small negative correlation between the “partial” category definitions and the matching rubric scores (-0.238 Kendall’s Tau correlation coefficient, p= 0.03). The correlations tests might not show significant results because of problems with instructional implementation of the crux concept, low sample numbers, and/or the concept itself may not be valid.

One interesting difference between crux vocabulary selected by the experts versus the students is the nature of expert-selected terms – the expert-selected terms were “easy.” They were terms that students might comfortably assume they knew, such as a-priori power analysis, aspirin dose, bleeding event, and cardiovascular events. Conversely, the terms selected by the largest number of students tend to be highly specialized terms such as Kaplan Meier Curve (23 students), Charlson Co-Morbidity
Index (22 students), conditional linear regression (16 students), Cox Proportional Hazard Model (18 students), and ICD9 Codes (13 students, also a crux term).

Several factors might have contributed to the student selection of these terms (and avoidance of the other terms) including the tendency to associate the concept of important vocabulary to new or unfamiliar concepts, the recognition that appropriate/inappropriate statistical tests can impact study results along with a general lack of knowledge about statistics, and biases or misconceptions that because a term is already familiar, they underestimate its relative importance in a study.

Conclusion Two

The second conclusion is that prior to reading the RCT, students’ initial clinical conclusions more closely resembled the experts’, potentially indicating an unbalanced influence of the RCT, either from RCT bias or a lack of skills transfer in evaluating the RCT.

**Student ability to reach a clinical conclusion.** There are two notable changes in the students’ clinical conclusion between the cohort/case-control decision (both studies had similar research conclusions) and the decision that also included the RCT (which had a different research conclusion from the cohort and case control studies). The first change in students is their clinical conclusion itself (some shifts were statistically significant others were not) and the second is a change in the level of student confidence around their decision (which was statistically significant).

The RCT appeared to influence the students’ clinical conclusion by shifting the number of students who first gave a more “conservative” clinical conclusion to use an \( H_2 \) antagonist instead of omeprazole (which was closer to the expert clinical conclusion) to
using omeprazole with clopidogrel despite the potential interactions shown by the cohort and case-control studies. Initially, 47% of the students chose an H$_2$ antagonist as their primary clinical option, which decreased to 33% after reading the RCT (not a statistically significant change). In the first clinical conclusion, 22% of the students chose to pair omeprazole with clopidogrel, which increased to 36% after reading the RCT (not a statistically significant change).

In addition, the percentage of students listing multiple or combination clinical options, such as considering the use of an alternative PPI like pantoprazole and/or an H$_2$ antagonist depending on the situation, decreased from 43% to 23% (this change is statistically significant, p=0.0232). Thus, students made a more singular clinical conclusion after the RCT. The students’ initial clinical conclusion more closely resembled the experts’ clinical conclusion.

The second change that was observed is that even though there were shifts in the clinical conclusion, student confidence in their decision increased. This is despite the fact that the results of the RCT conflicted with the results of the cohort and case-control studies, which one would think would decrease confidence.

After reading and evaluating the cohort and case-control studies, 42% of the students indicated they were confident in their decision. With the addition of the RCT, 62% of the students indicated they were “confident” in their decision, again despite the shifts in the clinical conclusion and the conflicting results of the RCT with the cohort and case-control studies (this is a statistically significant change, p=0.033). The changes in the clinical conclusion and changes in student confidence in their decision following the addition of the RCT potentially indicates bias towards RCT at the expense of critical
evaluation and/or a lack a skills transfer to evaluating the RCT (the RCT was a complex, challenging study to evaluate).

**Conclusion Three**

The third conclusion is the instructional model needs further development by adding explicit instructional scaffolding around the Medical Literature Evaluation (MLE) Rubric, vocabulary, and directly addressing student preconceptions/biases.

Four instructional scaffolds were implemented in this activity, the MLE Rubric, crux vocabulary, small group work, and whole class instruction. Each scaffold was associated with a theoretical framework and evidence-based instructional practices. Constructivism, adult learning theory, and science literacy theory are the frameworks that guided the evidence-based instructional approaches incorporated into the model for this activity and associated research.

Scaffolding is "the systematic sequencing of prompted content, materials, tasks, and teacher and peer support to optimize" independent learning (Dickson et al., 1993). Research supports the effectiveness of scaffolding, but it also shows that some approaches support learning better or differently than others (Davis & Linn, 2000; Chen & Bradshaw, 2007; Brunvard & Fishman, 2007), which is important to consider here.

Despite the previously successful tests of the instructional model, there appear to be limitations with such a complex scenario as the one associated with this activity. The instructional model needed additional explicit scaffolding around the MLE Rubric, vocabulary, and directly addressing student preconceptions/biases. In addition to making adjustments to the model itself, it also needs to be placed in a larger curricular framework to meaningfully advance the students in EBM. This section of the chapter elaborates on
the strengths and limitations for each instructional element of the model and provides suggestions for modifications to improve the model.

**The MLE Rubric.** The MLE Rubric is based upon existing EBM editorial guidelines (CONSORT, STROBE, and TREND) and other literature review instruments (MERGE and the Scottish Intercollegiate Guidelines Network Methodology Checklists). It has gone through numerous edits over several years of testing and implementation and has been reviewed by seven different experts in the field.

The rubric applies the constructivist and science literacy reading comprehension concepts of making the text organization explicit (by following the standard outline of a scientific paper) and defining clinical research-based vocabulary (e.g., allocation concealment, ancillary analyses, adverse events) and it applies the adult learning theory of self-direction in student learning.

The rubric is organized following the typical text structure of clinical research - Introduction, Study Design, Statistics, Results/Conclusions - and highlights elements that need to be contained/described in each section of the study. It also identifies what constitutes higher quality versus lower quality research and reporting (as identified by the EBM community). The intent is that the rubric can serve as an independent tool to guide students through the literature evaluation process, teaching them how to evaluate literature through descriptions and scales.

Although the MLE Rubric can be a useful tool, and students indicated that it contributed to their learning (39% indicated a moderate contribution and 38% indicated a high contribution), the data show its limitations as an independent training tool, at least in the early stages of student learning. The rubric may need to be introduced and reviewed
more explicitly with students using a variety of medical studies before attempting to
apply it in clinical conclusions using multiple studies of varying quality.

The data from this research show a difference between the expert MLE Rubric
Scores and student scores on the whole and within individual sections of the rubric,
particularly for the case-control and RCT studies (The RCT is the study with the most
differences between the student and expert scores). Unfortunately, the exact cause of
these deficiencies cannot be discerned from this research. For example, were students still
too lacking in background experience for this rubric to work? Was there a limitation to
the rubric itself for teaching (too complex or not enough details)? Or was the instruction
not explicit enough around using the rubric and applying it to different studies?

The only way to adequately address these questions would be to conduct
additional research targeting just the MLE Rubric (described later). My initial response to
these questions is that the students lacked appropriate literature evaluation background
experience and recent, explicit instruction around the rubric (recent is emphasized
because the students were introduced to the simpler version of the rubric for RCTs in
their second year and information and skills can be lost over time without practice, plus
this rubric addresses more than one study type).

In order to better design instruction, it is important to identify if there are strong
and weak areas for the students relative to the concepts within the rubric/evaluating
studies. Comparing student versus expert responses for the MLE Rubric subsections will
provide insight into this.

The first analysis identified the students’ strengths by looking at the study
elements where greater than 60% of the students matched the expert rubric score and
where students did well for all three study types. (60% was chosen because it is a common cut off point for “failure” and if more than 40% of the students are having trouble with an item, it is a high enough number that instructors need to address the item further.) This analysis shows that there were no study elements students did well in for all three studies. There are three examples where students matched the expert scores (>60%) in two of the three studies: control/comparison group (cohort and RCT), interim analysis (cohort and case-control), and outcomes and estimation (cohort and RCT). The background section came close for two of the three studies (the cohort and case-control).

The rubric elements where less than 50% of the students matched the expert scores are the weakest areas for the students. These areas would need deep and explicit explorations, particularly around the parts of the rubric where students have trouble with two or more of the studies. Thus, additional student instruction should consider how students responded compared to the experts.

Because there were so many rubric elements where the students did not match the expert score, it may indicate that the challenges in evaluating medical literature are considerably greater when mixing study types and study qualities. Even though there are many shared elements across the three study types, the transferability of literature analysis skills between studies appears to be limited.

Thus, the general indication is that students need more training on the elements of medical studies and identifying elements that are present, missing, or are of poorer/higher quality.

**The crux vocabulary.** Vocabulary in medical literature is complex and can change from study to study. Research shows that a reader’s background knowledge, the
text structure, text organization and coherence, and vocabulary are all important factors in reading comprehension and the level and frequency of new terms and concepts in the text also impact reading comprehension. (Kintsch, 1988, 1992). Although the meaning of new words in more general reading situations can often be derived from the surrounding text (Mezynski, 1983), this might not be true when reading scientific/medical literature that is vocabulary intensive and conceptually abstract. In vocabulary-dependent situations, research shows that a student’s vocabulary size and depth can impact reading performance—the larger the vocabulary repertoire the better the reading performance (Quian, 2002).

Because of the importance of vocabulary in reading comprehension and the amount of it in medical literature, a vocabulary-related scaffold was built into the instructional model. The initial version of the vocabulary scaffold used in the pilot study had students self-identify their own knowledge gaps by identifying and defining four to five vocabulary words and concepts with which they were not familiar or which they thought they needed to understand better to be able to evaluate the study. The original scaffold design addressed vocabulary challenges to enhance overall text comprehension and used meta-cognition, through the act of self-identifying unknown terms, to enhance learning.

The vocabulary exercise for this activity differed from the initial version in several ways. Rather than having students identify knowledge gaps and define words, this vocabulary exercise asked students to identify certain types of words and describe why they were critical for being able to evaluate the quality of the study. To do this, students must know the meaning of the word and be able to apply that meaning in the context of
the study design. This is a new concept I call “crux vocabulary,” which is defined as “key terms, concepts and approaches in medical studies that play a critical role in defining the quality of a study.” The crux vocabulary approach was developed for two reasons. One, to keep the students from becoming mires in the vocabulary, which can be extensive with multiple studies. Two, to help students make a cognitive leap differentiating “important” concepts from “critical” concepts.

The example of crux vocabulary provided to students in their instructions was:

”…in the red yeast rice study you read last year, the Food Frequency Questionnaire (FFQ) would be considered crux vocabulary” because you would need to investigate that term in-depth (beyond the information provided in the journal article) to recognize how that knowledge impacts one’s ability to evaluate the study. Although the survey was “validated,” as mentioned in the article, the original FFQ validation-publication stated that “the multiple R-squared using all factors to predict percent of calories from fat was 0.47” (0.47 is a low correlation). This is important because the FFQ played a large role in eliminating diet as a confounding factor in the study, but the because the FFQ R-squared was so low for determining calories from fat, it raises question about the reliability of the instrument for how it was used in the red yeast rice study. Thus, the conclusion the authors made that
changes in diet did not contribute to drops in cholesterol may be called into question.”

Three assumptions were made by asking the students to identify crux vocabulary. One was that the concept of crux vocabulary is valid. Second was that the expert identification of the crux vocabulary is valid. The third assumption was the students came with a certain level of background knowledge - they were now third-year students, they had studied cardiovascular pharmacotherapy, they had previously evaluated literature and they could therefore make the cognitive leap from basic vocabulary concepts to crux vocabulary.

The third assumption about students’ ability/background knowledge appears to have been incorrect since students had trouble both identifying the crux vocabulary (as identified by the experts) and providing “reasoning” level descriptions (as opposed to the lower-order descriptions of the term they selected). The experts identified 11 crux vocabulary, five of which were matched by no more than 13 students. Additionally, only 18% of the descriptions provided by students for any term they chose (be it crux or not) met the “reason-level” criteria students were asked to provide.

Because the data collected for this dissertation raise questions about the crux concept and/or its value, some additional research is needed. If valid, instructional investigations would be needed to determine how to better use the crux vocabulary in instruction and its relative contribution to student learning compared to the MLE Rubric and the other scaffolds.

As a starting place, some modifications to the instructional approach would be needed to enhance learning. Rather than ask students to identify crux words, students
might need to begin with a list of crux words as a stepping stone to developing an understanding of the concept and thinking like an expert. By providing the word list for the students, it would not only train the students about the words and save time, it could eliminate word choice bias, which is a bias against words that appear common to them even though they may be crux. In this case, students would intentionally avoid something they think they already know about, even if they do not fully understand the importance of the word to determining the quality of the study (e.g., cardiovascular events).

In addition to giving students the list of words, there would need to be explicit instructions to provide both definitions and the description of what makes the word “crux.” By requesting both, it would make a clear distinction for the students that a definition and the reason are not the same.

**The small and large group discussions.** The remaining scaffolds in the instructional model are the instructor-led large class discussion around the MLE Rubric, articles, and crux vocabulary and the small group peer discussion about the crux vocabulary and RCT.

The whole-class discussions were facilitated by one of the experts to review the crux vocabulary and key points about the MLE Rubric scores for the cohort and case-control studies to address questions and share the expert-identified crux words and reasons why they are crux words. As part of the class discussion, the expert asked students to share their scores for certain sections of the rubric and explain why they gave that score and he encouraged students to share and reflect upon their crux vocabulary. Then he shared the expert-identified crux vocabulary and why a term was crux and highlighted key points on the rubric for each of the three studies.
Students found the large class discussions about the MLE Rubric and articles to contribute the most to their learning (56% of the students rated it as having a high contribution to their learning and 33% a moderate contribution). Thirty-three percent of the students found the small-group peer discussion about the RCT to have a high contribution to their learning and 43% a moderate contribution. The class discussion around the crux vocabulary was less helpful according to students with just 5% rating it as having a high contribution to their learning, 38% a moderate contribution and 46% contributing a little bit to their learning.

Because there were two different parts to the larger class discussion, the cohort/case-control discussion and the RCT discussion, I cannot discern whether students felt part one or part two contributed differently to their learning. Students completed the MLE Rubric for the RCT after the large class discussion for part one (where the students learned about the expert scores for the cohort and case-control studies, why they were scored in a particular way, and crux vocabulary), but they learned about the major problems with the RCT and the expert clinical conclusion after part two. Those revelations about the RCT expert score and clinical conclusion might have impacted the students more because that was the point at which the students’ own RCT bias and different clinical conclusion compared to the experts was revealed, but this, again, cannot be discerned from the data.

Although students’ perceptions of their own learning can be a useful indirect measure, the student responses on the MLE Rubric for the RCT and the final clinical conclusion were direct measures of changes in student ability as a result of the expert guidance and class discussions about the cohort and case-control studies and the crux
vocabulary. With respect to the MLE Rubric, although the expert score and the average student score for the RCT are the closest of the three, with only a 4% spread (the consensus expert score was 73% and the mean student score was 77%), in this case the final MLE score is not the best measure to assess student learning compared to the experts, because the differences in the scores that produced the averages cannot be discerned. Instead one needs to look at the individual MLE Rubric fields.

There are several sources of data that show divergence of data of student responses from expert responses. When looking at the data from chapter four, the median scores for each section of the rubric for the RCT there are 16 fields where the students’ score has at least a 0.7-point difference compared to the experts’ score. This is compared to seven fields for the cohort and nine for the case-control. In addition, three of the RCT fields have a two point difference between the students’ median score and the expert score. There are no fields in the cohort or case-control rubrics with a two point difference. In all cases with the two point difference, the students gave a higher score (a score of two or three compared to the zero or one given by the experts).

Not only is a two point score difference large, the fields that have a two point difference are critical for the analysis of this particular study. For example, the median score given by the students for the objectives section for the RCT was three and the experts gave it a one. The experts assigned a low score because the study objectives broadly mentioned that it was a safety and efficacy study primarily around gastrointestinal (GI) outcomes. Mention of cardiovascular outcomes in the study objectives was very vague and came across as almost an afterthought. However, despite the original study design’s GI orientation, the paper ended up being a discussion about
cardiovascular effects (or lack thereof). Thus, there was a mismatch between the original objectives of the study and the conclusions the authors ended up making. Students did not catch this mismatch, they only saw that there were objectives present (despite seeing much higher quality cardiovascular objectives in the case-control and cohort studies).

In the sample size power calculation of the RCT the experts gave the study a zero and the student median score was two. Again, the students missed the fact that that the power calculation presented in the study was for gastrointestinal effects, but the focus of the analysis and the paper ended up being on cardiovascular effects. There was no power analysis performed for cardiovascular effects and, therefore, no way to know if the sample size upon which the authors made their conclusions was adequate. Although a few students caught the problem, most apparently assumed that the sample size was adequate for the all of the outcomes.

The third field where there was a two point difference was the interim analysis. The median student score was two and the expert score was a zero. This was because no interim analysis actually occurred due to the study ending prematurely from a loss of funding. This early termination potentially affects the appropriate timing to be able to make judgments about cardiovascular endpoints versus GI endpoints; the timing was reasonable for the GI but not the cardiovascular effects. If the study had just focused on the GI effects and was used in the field for that purpose then there might not be an issue with the study; however, the study shifted its focus to the cardiovascular endpoints and did not have an adjusted study design for cardiovascular effects. By reading the RCT, it also shifted the clinical conclusions of the students, moving them from a cardiovascularly conservative position based upon similar results from two studies, to a decision to co-
prescribe clopidogrel and omeprazole based upon one flawed “gold standard” study published in a world-class journal.

Because of the randomized and controlled study design, understandably RCTs are placed near the top of the hierarchy of evidence pyramid, below meta-analyses but above cohort and case control studies. This has resulted in a phenomenon called RCT bias, in that RCTs are often looked at almost exclusively in EBM decisions, excluding the outcomes from other study designs, such as case control or cohort (often referred to as observational studies) (Ashcroft, 2004; Colditz, 2010). Research is showing that other study designs may have their place in EBM because different study designs investigate different constructs. For example, RCTs might examine a new therapy versus an alternative therapy but an observational study might examine the impact of a therapy in those who received it versus those who never received that therapy (Colditz, 2010). Ethically an RCT may not be able to investigate the difference between a treatment versus no treatment or combinations of treatments that may cause harm. In addition, observational studies can look at very large populations compared to RCTs, thus potentially seeing trends that might not show up in an RCT simply because of the study size.

Despite those facts, RCTs are held as the gold standard study design and are often promoted as the study design to be used for decision-making. As a general principle, using RCTs as a default study design for decision-making is a reasonable concept, however, the questions arise: Are students’ critical thinking and analysis abilities potentially compromised by heavily emphasizing the relative importance of RCTs compared to the other study designs?; and Does that emphasis introduce a, perhaps
unintended, bias into their training?

Preconceptions, which include bias and misunderstandings/fact problems, have been shown to have a major impact on student learning. Posner, Strike, Hewson, and Gertzog (1982) proposed the Conceptual Change theory, which states that learning science is adversely impacted by incompatible beliefs and misconceptions held by students and that those misconceptions must be explicitly critically evaluated by students and revised for accurate learning to occur. In the research-based publication “How People Learn: Bridging Research and Practice” (National Research Council, 2006), it emphasizes several key findings in science education research one of which is:

Students come to the classroom with preconceptions about how the world works. If their initial understanding is not engaged, they may fail to grasp the new concepts and information that are taught, or they may learn them for purposes of the test but revert to their preconceptions outside the classroom (p. 10).

The National Research Council elaborates upon that statement indicating that the initial understandings students develop can “have a powerful effect on the integration of new concepts and information.” Accurate understandings can serve as a foundation for new knowledge, however inaccurate understandings, biases and stereotypes can be very difficult to overcome in instruction and learning (National Research Council, 2000). The results of this dissertation appear to support the premise that bias or preconceptions can adversely impact student learning.

Thus, in considering the instructional model to teach students how to critically
evaluate medical literature and reach a clinical conclusion, another scaffolding element needs to be added. That scaffold is directly addressing, in a clear and structured manner in the instruction, the preconceptions/biases that students bring with them related to RCTs.

Research into overcoming pre-conceptions in order to facilitate learning science show several things: (1) more prior knowledge reduces misconceptions (Braasch & Goldman, 2010); (2) students that view science as “dynamic and tentative” rather than “static and fixed” have fewer misconceptions (Linn & Songer, 1993); and (3) misconceptions can change through experience and providing instruction that set’s up an experience of cognitive dissonance followed by a logical process for constructing the new knowledge has been shown to be effective (Enderle, Smith, & Southerland, 2009).

When considering applying these instructional concepts to the model, perhaps the best type of scaffolds to help overcome this bias would be to review the MLE Rubric in detail with the students prior to the activity (building prior knowledge) and simultaneously discuss study design strengths and weaknesses, variability in the quality of all types of studies, and the potential for RCT bias and cases where other study designs may be useful (introducing the concept that science is dynamic).

A next step would have the students to complete the activity and the instructor conduct a class poll and discussion around the student’s article rankings and clinical conclusion. The instructor later reveals the expert’s evaluations and clinical conclusion, and, if bias were present amongst the students, reveals that bias (setting up cognitive dissonance). If bias appears to be present, students can work in small groups and discuss the differences between their responses and the expert responses, reflecting upon how the
presence of bias may have impacted their clinical conclusion and potentially patient health (logical process for constructing new knowledge).

Implications for the Field

This research has multiple, important potential implications in pharmacy and medical education, as well as more broadly in EBM. In pharmacy and medical education, this research provides more details about what may be needed to really teach students how to practice EBM. Medical education research shows there is already a deficit in EBM instruction in the medical curriculum (Green, 2000), and this research demonstrates that, even with a comparatively strong presence of EBM in the curriculum, it still may not be enough to ensure students are able to make informed medical decisions from different types of medical literature of varying quality and with different conclusions. Thus, it appears specific instructional components and approaches need to be present in EBM instruction to maximize student learning.

This research also demonstrates why teaching EBM is important, since a single flawed study can potentially influence students’ clinical conclusions, even when considered in the context of multiple other studies with conclusions. The question exists regarding the general ability of medical practitioners to evaluate medical literature and reach sound clinical conclusions, since their training may be limited.

More broadly, this research can provide insight into how to teach graduate students in other science disciplines how to evaluate literature in their fields and use multiple studies to reach conclusions. Lastly, although this study relates to a very specific field of education and revealed how RCT bias can impact clinical conclusions, it could
have a broader interest to those investigating the influence of bias or preconceptions on student learning.

**Future Research**

There are several areas for future research related to this work including research around the MLE Rubric, crux vocabulary, teaching students to reach a clinical conclusion, and investigating the potential influence of RCT bias and the application of this instruction in an applied setting, such as during rotations.

With respect to the MLE Rubric, additional studies could investigate instrument validity and design (e.g. weighting the study elements). The rubric design can also be investigated to maximize the rubric’s instructional use, such as varying the level and type of detail in the instrument.

Other areas for future research include investigating the validity of the crux vocabulary concept and testing changes to the instructional approach for crux vocabulary and/or comparison studies with respect to the contribution of vocabulary to being able to evaluate a study.

A third area for future research is related to instruction around teaching students to reach sound clinical conclusions. It could address questions about the effects of different studies of varying quality on changes to student’s clinical conclusions, as well as how to address potential misconceptions/bias students may have about RCTs.

Lastly, it would be beneficial to see how the rubric and the skills learned from the activity are applied, if at all, in a clinical setting, like during rotations.

**Conclusion**

There are three main conclusions as a result of this research:
• Conclusion One - Third-year doctor of pharmacy students at the COP have a limited ability to evaluate medical literature of varying qualities and types and conflicting conclusions;

• Conclusion Two - Prior to reading the RCT, students’ initial clinical conclusions more closely resembled the experts’, potentially indicating an unbalanced influence of the RCT, either from RCT bias or a lack of skills transfer in evaluating the RCT; and

• Conclusion Three – The instructional model needs further development by adding explicit instructional scaffolding around the Medical Literature Evaluation (MLE) Rubric, vocabulary, and directly addressing student preconceptions/biases.

These discoveries show that more explicit scaffolding needs to occur in this activity to better support student learning, but ideally it also needs to occur in a progression throughout the curriculum. Three areas need particular focus. These areas are (1) more training with the MLE Rubric including detailed review of all of its components, (2) more direct interaction with different types of studies and evaluating those studies, and (3) addressing potential RCT bias. Another area that should also be explored, both in terms of validity and instruction, is the crux vocabulary.

Tables 25 and 26 show the recommended changes to the activity presented in this study, as well as a recommended curricular progression to get students to the point of being able to do and fully benefit from this activity.
Table 25

Recommended Changes to the Activity

<table>
<thead>
<tr>
<th>Suggested Revised Activity Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Detailed MLE rubric review (new)</strong> – Originally it was thought that the MLE Rubric could be used as an independent/stand alone tool, that it, by itself, could “train” the students. This research did not support that premise, thus the tool and the study evaluation concepts imbedded within it require explicit instruction/training. Ideally this would be done with specific, short examples from medical studies. This would also be the time to highlight the strengths and weaknesses of the different study designs and begin to address RCT bias.</td>
</tr>
<tr>
<td><strong>Crux vocabulary review (new)</strong> – (Assuming the validity of the crux vocabulary concept) Students would be more formally introduced to the concept of crux vocabulary, reviewing several examples. A new scaffold would be given to the students, which is the expert list of the crux vocabulary associated with the medical studies used in the activity. Students would be instructed to look up the definition of each word AND they must provide a detailed description of why the term is considered crux. This exercise could be subdivided by small groups, with students dividing the words amongst the group and sharing and discussing their answers.</td>
</tr>
<tr>
<td><strong>Students independently and individually evaluate the cohort and case control studies (original design)</strong> - This can be completed outside of class and students can bring notes and rubric scores to class in preparation for the small and large group discussions.</td>
</tr>
<tr>
<td><strong>Small and large group discussions (original design and new)</strong> – Students share their scores and discuss scoring for the cohort and case control studies and discuss their clinical conclusion. In small groups students discuss the crux vocabulary in detail (new).</td>
</tr>
<tr>
<td><strong>Students individually evaluate the RCT and generate a clinical conclusion based upon the three studies (original design)</strong> – Students read and individually evaluate the RCT using the MLE Rubric.</td>
</tr>
<tr>
<td><strong>Small and large group discussions (original design and new)</strong> - Students share their scores and discuss scoring for the RCT and discuss their clinical conclusion. The instructor conducts a class poll and discussion around the student’s article rankings and clinical conclusion, reveal the expert’s evaluations and clinical conclusion, and determine if bias was present amongst the students. If so, then have the students work in small groups and discuss the differences between their responses and the expert responses, reflecting upon how the presence of bias may have impacted patient health. The discussion would also include an analysis of differences between individual responses and expert responses.</td>
</tr>
</tbody>
</table>
Table 26

*Suggested Curricular Progression to Support Student Learning*

<table>
<thead>
<tr>
<th>Suggested Curricular Progression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year 2, semester 1</strong></td>
</tr>
<tr>
<td>– RCT evaluation activity 1. (Currently in the curriculum)</td>
</tr>
<tr>
<td>– RCT evaluation activity 2 – students use the MLE Rubric to evaluate 2 RCTs of differing quality on the same topic. (New)</td>
</tr>
<tr>
<td><strong>Year 2, semester 2</strong></td>
</tr>
<tr>
<td>– Students use the MLE rubric to evaluate a cohort study and a case control study. They can be on different topics or matching topics. This would be an instructor-lead activity. (New)</td>
</tr>
<tr>
<td>– Students complete the group literature review/research paper in the informatics and research course, but perhaps expanded to include the independent use of the MLE Rubric. (Currently in the curriculum, but modified).</td>
</tr>
<tr>
<td>– Students would utilize the MLE Rubric in the existing lab journal club. (Currently in the curriculum, but modified).</td>
</tr>
<tr>
<td><strong>Year 3, semester 1</strong></td>
</tr>
<tr>
<td>– Students complete the MLE activity used in this research. (Currently in the curriculum)</td>
</tr>
<tr>
<td><strong>Year 3, semester 2</strong></td>
</tr>
<tr>
<td>– Students complete the independent literature review and clinical recommendation activity. (Currently in the curriculum.)</td>
</tr>
</tbody>
</table>
Appendices
Appendix A: Clinical Literature Evaluation (CLE) Rubric

**Study title:**

**Study goal/objective/hypothesis:**

**INSTRUCTIONS:** As you read through the article, use this rubric to help you evaluate the article’s quality. Evaluation criteria are divided by general article sections: Introduction/General; Study Design; Statistics; Results/Conclusions. Identify the appropriate description that most closely resembles the study, write the corresponding score in the column labeled “Score,” and write notes to help you recall why you assigned a particular score for each section. The shaded columns to the far left are excerpts from the CONSORT (Consolidated Standards of Reporting Trials) Statement and other additional information to further assist your evaluation.

**Introduction Section**

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Background:</strong> Scientific background and explanation of rationale.</td>
<td>Background information is not included.</td>
<td>Background information is included, but some important information is left out.</td>
<td>Most of the relevant or needed background information is included.</td>
<td>Background information is thorough.</td>
<td></td>
</tr>
<tr>
<td><strong>Objectives:</strong> Specific objectives and hypotheses (questions the study was designed to answer).</td>
<td>Does not include objectives or hypotheses.</td>
<td>Includes objectives or hypotheses but are not clearly written or obvious or are difficult to measure.</td>
<td>Includes measurable objectives or hypothesis but they could be more specific or clearer.</td>
<td>Includes clear, specific and measurable objectives or hypotheses.</td>
<td></td>
</tr>
</tbody>
</table>

**Introduction Section Score**

(total points possible 9)

**Notes:**

---


Dawn, et al. 2011
### Study Design/Methods Section

**CONSORT Statement**

**Description**

**Score**

<table>
<thead>
<tr>
<th><strong>Trial design:</strong> Brief description of the type of trial (eg, parallel, multi-arm parallel, crossover, cluster, factorial) and conceptual framework (superiority or non-inferiority).</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial design is not described.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial design description is missing very basic, important information.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial design is mostly described but is missing some key information.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial design is well described and includes statement of trial type and conceptual framework.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Blinding (masking):** Whether participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated. Prevents performance and ascertainment bias. Blinding is particularly important with subjective outcome measures.

**Description**

**Score**

| Study does not mention if it is blinded. | | | | |
| Study is not blinded or claims blinding but does not describe who was blinded, so it is difficult to determine type of blinding. | | | | |
| Single-blind study design; study describes who was blinded. | | | | |
| Double- or triple-blind study design; study describes who was blinded. | | | | |

**Example where no blinding is needed (at least for the patients):** comparing the pharmacokinetics of two dosage forms.

**Control:** Comparison group for the experiment. In clinical trials a placebo group is a common control, but other control groups may be helpful to minimize false negatives or positives. Active treatment may be a comparison drug. Ethical issues may impact use of control groups (eg, withholding treatment from cancer patients is considered unethical). Characteristics of control and intervention need to be similar (appearance, taste, smell, administration).

**Description**

**Score**

| Study does not have a control (eg, comparison group or placebo) | | | | |
| Study has a control, but it is not the best choice for a control or there are some problems with the control design (eg, problems with appearance, taste, administration), or control design is not well described. | | | | |
| Control for the study appears to be adequate based upon the description, but details are not explicit. | | | | |
| Study is very well controlled; or controlled as needed and details are explicit. | | | | |

**Eligibility criteria & setting:** Eligibility criteria (EC) for participants (eg, method of recruitment, nature and stage of the disease, exclusion criteria) and the setting and locations where data are collected (eg, country, city, multi-center, hospital clinic, whether setting might influence results – such as transportation issues).

**Description**

**Score**

| Study does not describe EC & setting; or EC are not applied equally to all study groups thereby introducing possible selection bias. | | | | |
| EC are briefly mentioned with little or no justification for the criteria. EC are on the weak side, allowing potential confounding factors. EC might not be applied equally to all study groups. Setting details are not provided. | | | | |
| EC are well described and eliminate the largest and most obvious confounding factors. They appear to be applied equally to all study groups. Justification of EC might be limited. Setting descriptions appear to be complete. | | | | |
| EC are well described, justified, and rigorous, eliminating multiple potential confounding factors. It is clear that the EC have been equally applied to the study groups. Setting is described in detail and carefully considered in the design. | | | | |

**Randomization:** Sequence generation – method used to generate the random allocation sequence (eg, random number table or generator), including details of any special features/restrictions (eg, blocking, stratification), and how code breaking is avoided; Simple randomization is often best but other restrictions may occur in studies with smaller numbers Allocation by alternation, hospital numbers or date of birth are NOT random.

**Description**

**Score**

| Original sample selection is not random, but could have been and study could benefit from it; or sample selection is not stated | | | | |
| Original sample selection is not random and is needed to better control the study. | | | | |
| Original sample is randomly selected from a small population. | | | | |
| Original sample is randomly selected from a large population. | | | | |
### Allocation ratio and concealment: Description of allocation ratio (intended number of participants for each comparison group, eg, 1:1) and method used to implement the random allocation sequence (eg., numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned. Allocation concealment prevents selection bias, third-party assignment is best.

<table>
<thead>
<tr>
<th>Treatments/group assignments are not randomized, a description of the “randomization” process is not provided, or randomization is claimed but the method is not truly random. The allocation ratio is not described.</th>
<th>Treatments/group assignments are not randomized, and study provides a reasonable and logical explanation. Allocation ratio is provided. If it is an unequal allocation (eg, 1:2) the rationale is not explained or rationale is questionable.</th>
<th>Treatments/group assignments are randomized and methods well described and acceptable relative to participant numbers; Allocation ratio is not well described or justified.</th>
<th>Treatments/group assignments are randomized and allocation ratio is provided; both are well described and justified relative to participant numbers; care was given to avoid bias through inadvertent code-breaking or too many strata.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implementation: Who enrolled participants, who generated the allocation sequence, who assigned the participants to their groups; and who administers intervention.</td>
<td>The same person(s) enrolled the participants, assigned to treatments, as well as implemented.</td>
<td>The study attempts to separate enrollment, allocation, and implementation; there is overlap in the people who allocate, enroll, and implement.</td>
<td>Implementation of randomization processes are described, some of the same people participated in more than one of the processes (eg, sequence generation and allocation concealment) but not also implementation.</td>
</tr>
<tr>
<td>Treatment plan/regimen is not described.</td>
<td>Treatment plan/regimen is described but contains some weaknesses.</td>
<td>Treatment plan was clearly described and appears to be appropriate.</td>
<td>Treatment plan/regimen was well explained, well justified, and logical.</td>
</tr>
</tbody>
</table>
**Study Design/Methods Section, cont’d**

<table>
<thead>
<tr>
<th>CONSORT Statement Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcomes:</strong> Clearly defined primary and secondary outcome measures (including defining objective and subjective measures) and, when applicable, any methods used to enhance the quality of the measurements (eg, multiple observations, training of assessors, use of “gold standards” or “state of the art” measures, whether instruments are validated)</td>
<td>Primary and secondary outcomes are not well described or differentiated and there are too many primary outcomes that may contribute to problems with multiplicity of analyses. Methods used and associated outcome measures are not well described and inappropriate to address objective(s)/hypotheses.</td>
<td>Primary outcomes are described, secondary outcomes mentioned but not well described. Methods used and associated outcome measures are described but lack key details and not all outcome measures seem to be appropriate to address objective(s)/hypotheses (eg, subjective measures are used when objective outcomes are available or lower quality subjective measures are used).</td>
<td>Primary and secondary outcomes are described, measures are defined and, for the most part, well-matched to objective(s)/hypotheses but could be more rigorous. There is a mix of objective, partially objective, and subjective measures. Partially objective or subjective measures are mostly appropriate but there may be a higher quality measure available.</td>
<td>Methods and primary and secondary outcomes and measures are well defined, thorough and appropriate to address all objectives/hypotheses. Objective measures are primarily used, use of partially objective and subjective measures are well justified.</td>
<td></td>
</tr>
<tr>
<td>a. <strong>Examples of objective measures:</strong> assessment of bone mineral density on multiple skeletal sites by DEXA; repeated assessments of pulmonary function.</td>
<td>The study has no quality enhancement measures (eg, multiple observations, training of assessors, or use of gold standard measures).</td>
<td>The study has very limited quality enhancement measures.</td>
<td>Quality enhancement measures are present but limited; they address key areas.</td>
<td>Quality enhancement measures are thorough and rigorous. Study uses gold standard measures.</td>
<td></td>
</tr>
<tr>
<td>b. <strong>Examples of partially-objective:</strong> abstracting ICD-9 codes from administrative database for the diagnosis of osteoporosis; Partially-objective: number of hospitalizations and unscheduled clinic visits from administrative database.</td>
<td>It appears there may have been a change in the study design but the authors are not explicit about the change and reasons for the change, or the change may introduce bias (eg, changing an end-point based on unblinded data without an independent data monitoring committee).</td>
<td>Changes to the initial study protocol or outcomes are described; justifications are not included, and it is unclear if bias may result from the change.</td>
<td>Changes to the initial study protocol or outcomes are described; justifications are included but limited and appear to not introduce bias.</td>
<td>Changes to the initial study protocol or outcomes are well described, well justified (eg, evidence from a systematic review suggests an end-point is not appropriate), and will likely not introduce bias. OR No changes were needed and did not occur, which was explicitly reported.</td>
<td></td>
</tr>
<tr>
<td>c. <strong>Examples of subjective measures:</strong> survey to ascertain if a patient has been diagnosed with osteoporosis; asking the patient to report asthma severity (eg., frequency of symptoms, use of rescue medications).</td>
<td>NOTE: Certain types of studies only have subjective or partially objective measures, this should be taken into consideration and noted below.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dawn, et al. 2011
### Statistics Section

<table>
<thead>
<tr>
<th>CONSORT Statement Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistical methods:</strong> Statistical methods used to compare groups for primary outcome(s); methods for additional analyses, such as subgroup analyses and adjusted analyses (eg, for multiple comparisons); distribution of data (eg, normal or skewed) are provided to be able to judge test choices. Confidence intervals (at least 95%) and p-values should be included; multiple observations per individual require advanced statistics and cannot be treated as independent data; subgroup analyses are clearly specified and use complementary subgroups and a test of interaction (simple comparison of P-values has high false positives), but additional subgroup analyses are often discouraged.</td>
<td>No statistics or inappropriate statistical tests/analyses were performed or the study does not name the tests used.</td>
<td>Minimal statistical tests/analyses were performed or minimal descriptions of the tests and why they were chosen is provided.</td>
<td>Adequate descriptions of the distribution of data and statistical tests/analyses are provided; appropriate tests performed.</td>
<td>Thorough descriptions of the statistical tests/analyses are provided and justified relative to the distribution of the data and appropriate tests were performed.</td>
<td>(total points possible 12)(one possible 3-point denominator deduction)</td>
</tr>
<tr>
<td><strong>Sample size (power calculations):</strong> How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules. Post-hoc calculations of statistical power has little merit, this needs to be expressed via confidence intervals.</td>
<td>Sample size/power calculations/justifications are not described.</td>
<td>Sample size/power calculations/justifications are described but important details are omitted.</td>
<td>Sample size calculations/justifications are described lacks minor details.</td>
<td>Sample size calculations/justifications are well described and seem reasonable.</td>
<td></td>
</tr>
<tr>
<td><strong>Interim Analyses:</strong> Interim analyses, or examining results as data accumulate, may occur to monitor an intervention and need to be reported. There are specific group sequential statistical methods are available to adjust for multiple analysis and they, along with stopping rules, should be pre-specified in the protocol.</td>
<td>Interim analyses occurred but no information is provided about how the data were analyzed; analyses were not pre-specified in the protocol; analyses were not appropriate (eg, using a p&lt;0.05 for multiple interim analyses) and can result in false positives/inaccurate results; stopping rules are unclear or not described.</td>
<td>Interim analyses were designed as part of the original protocol, but descriptions or justifications are not well described and it is unclear. If appropriate statistical methods were used, stopping rules are well defined.</td>
<td>Interim analyses were designed as part of the original protocol, but descriptions or justifications for interim analyses are less detailed but can infer from them. It appears that appropriate statistical methods and stopping rules are well defined.</td>
<td>Interim analyses were designed as part of the original protocol and use well described, well-justified and appropriate statistical methods. Stopping rules are well defined.</td>
<td></td>
</tr>
</tbody>
</table>

If no interim analysis occurred then select this box and remove 3 points from the denominator.

<table>
<thead>
<tr>
<th>Statistics Section Score</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dawn, et al. 2011</strong></td>
<td></td>
</tr>
<tr>
<td>Results/Conclusions Section</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>CONSORT Statement</strong></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>0</td>
</tr>
<tr>
<td><strong>Participant flow:</strong> Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome. Describe protocol deviations from study as planned (including stopping the trial), together with reasons.</td>
<td>Participant flow information is excluded or very poorly described. No explanations are provided for participant dropout and protocol deviations.</td>
</tr>
<tr>
<td><strong>Key Data:</strong> Recruitment: Dates defining the periods of recruitment and follow-up. Baseline data: Baseline demographic and clinical characteristics for each group.</td>
<td>Key data are omitted in the text and not presented in tables or graphically.</td>
</tr>
<tr>
<td><strong>Numbers analyzed:</strong> Number of participants (denominator) in each group included in each analysis and whether the analysis was by “intention-to-treat.” State the results in absolute numbers when feasible (eg, 10/20, not 50%).</td>
<td>Did not include the number of participants completing study.</td>
</tr>
<tr>
<td><strong>Outcomes and estimation:</strong> For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (eg, 95% confidence interval). Binary outcomes (eg, death or oxygen dependence) should be presented as both absolute (risk difference) and relative effect (risk ratio/relative risk or odds ratio) sizes.</td>
<td>Results are provided for only the primary outcome and does not provide confidence interval.</td>
</tr>
<tr>
<td><strong>Ancillary analyses:</strong> Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory. One analysis that should be included is the Intent to Treat (ITT) analysis which compares patients in the groups to which they were randomly assigned (eg, based on the initial treat intent), the treatment they actually received, and withdrawal or protocol deviation (thus includes ALL patients, even those who were later excluded or dropped out) (Hollis &amp; Campbell, 1999) If an ITT is not performed it is well justified and describes other analyses (eg, efficacy or per-protocol analysis)(Greenhalgh, 2006). Multiple analyses of same data create a risk for false positives. Sub-group analyses should be avoided (see statistics section). Pre-specification in trial protocol is more reliable.</td>
<td>Ancillary analyses are not performed or described.</td>
</tr>
</tbody>
</table>
### Adverse events/harms:

All important adverse events or side effects/harms in each intervention group.

Consider whether adverse events were detected using passive or active processes.

Adverse events need to be well defined (What are they specifically looking for and how is it measured?)

Consider whether adverse events are from treatment or the disease/condition.

Look for reporting of and consider the rate of adverse events in each group.

- **Adverse events are NOT clearly described.**
- **Adverse events are briefly described and explanations for potential impacts on the study are limited and have some questionable logic, but may slide by.**
- **Adverse events are described, but some details are excluded (eg, why specific measures were chosen). A logical explanation is provided regarding potential impacts to the study.**
- **Adverse events are clearly described and thoroughly addressed (including how adverse events and side effects were detected and measured, why measured were used, used thorough and appropriate measures, and logical explanations of how the study or conclusions may be impacted).**

### Interpretation:

Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision, and the dangers associated with multiplicity of analyses and outcomes.

- **Bias from conflict of interest is clearly present and not disclosed by authors.**
- **Bias or potential conflict of interest is present and disclosed by authors, but could impact study.**
- **Bias or potential conflict of interest is present and disclosed by authors, but appears to have minimal impact on study.**
- **Bias or conflict of interest is minimal and is not likely to influence results.**

### Potential Conflict of Interest:

Did the authors declare potential conflicts of interest? Are they affiliated with the company that manufactures the drug or device being tested?

### Generalizability:

Generalizability (external validity) of the trial findings.

### Overall evidence:

General interpretations of the results in the context of current evidence.

- **Obvious study limitations are not reported (eg, potential bias and confounding).**
- **Study limitations are mentioned but unclear or some limitations are overlooked.**
- **Study limitations are adequately reported, external validity is discussed, but their potential impact on study results are not well described.**
- **Study limitations are reported, external validity is discussed, potential bias and confounding are acknowledged, and their potential impact on the study results are well described.**

### Discussion/interpretation of the results

- **Discussion/interpretation of the results are inconsistent with data and misleading.**
- **Some of the discussion/interpretations of the results are consistent with the data, but some inconsistencies or overstatements exist.**
- **Most of the discussions/interpretations of the results appear to be consistent with the data.**
- **Discussions/interpretations of the results are very consistent with the data.**

### Study limitations

- **Obvious study limitations are not reported (eg, potential bias and confounding).**
- **Study limitations are mentioned but unclear or some limitations are overlooked.**
- **Study limitations are adequately reported, external validity is discussed, but their potential impact on study results are not well described.**
- **Study limitations are reported, external validity is discussed, potential bias and confounding are acknowledged, and their potential impact on the study results are well described.**

---

Dawn, et al. 2011
Appendix B

Characterization Survey

**Question 1:** Please enter your assigned study ID number.

**Question 2:** How familiar are you with evidence-based medicine? (somewhat, moderately, very)

**Question 3:** How important should evidence-based medicine be in pharmacy practice? (not at all, moderately, very)

**Question 4:** How often do you read primary scientific literature (i.e. full publications of the original research) from research journals (this includes medical and bench research)? (not at all, rarely, a few times a year, a few times per month, a few times per week)

**Question 5:** Approximately how many primary scientific literature publications have you read (this can include any scientific or medical subject) (fewer than 5, 5-10, 11-20, 21-40, >40)

**Question 6:** Have you personally participated in conducting scientific research? If yes, select all that apply. (no; yes, as a high school student; yes, as a pre-pharmacy student; yes, in a previous job)

**Question 7:** Have you been an author or co-author on a scientific paper? (yes, no)

**Question 8:** Rate yourself on the following: (not at all, a little, mostly, definitely)

- I know what specific things to look for to determine the quality of a research journal;
- I know what specific things to look for to determine the quality of a single, published research study;
- I am confident in my ability to look up or obtain information to help my understanding of a published research study;
- I can confidently read and understand a published research study inside of my field of interest;
- I can confidently read and understand a published research study outside of my field of interest;
- I can confidently evaluate a randomized controlled trial (RCT);
- I can confidently evaluate a cohort study;
- I can confidently evaluate a case control study;
- I can confidently make a clinical decision using multiple individual studies (excludes reviews and clinical guidelines).
Appendix C: Pre-lab MLE Activity Instructions
Prior to Lab
Medical Literature Evaluation and Reaching a Clinical Conclusion Activity
Instructions

Pre-lab work 5 Points

ALL STUDENTS: In preparation for your lab the week of January 23, 2012 you must complete the following prior to coming to class (NOTE: There will be no other homework associated with this activity, it is all done up-front):

1. Read the two articles (posted in WebCT under week 2 lab):

2. Complete the Pre-lab Assignments by doing the following (all of these items are contained in the Word document “MLE Pre-lab Assignment” that is being emailed to you. PDFs are also in WebCT, but please complete the assignment electronically in Word – note the PDFs of the “MLE Pre-lab Assignment” document were split into two documents when converted.):
   a. Fill out the Score Sheet and MLE Rubric Notes.
   b. Complete the Crux Vocabulary assignment.
   c. Email your completed work before your lab date to the following email address to receive credit for this work:
      i. Email to: (email removed)
   d. Bring your Crux Vocabulary words with you to class for the discussion.

3. Clinical Decision: Go to the website listed below and submit the clinical decision you would make based upon the case control and cohort studies you read. You must enter your assigned ID number to receive credit: (website removed)

4. OPTIONAL: Complete the Comparing the Studies worksheet for the two studies.
   a. A Word document of this has been emailed to you (a PDF of this document is in WebCT and a Word version is also in SharePoint)
   b. This is a useful tool to help you keep track of the different features of each study. It is easy to fill out as you are reading the article and you won’t have to keep returning to the study to find this important information.

RESEARCH PARTICIPANTS: Prior to coming to your lab the week of January 23, 2012 please complete the brief (10-min) 10-item Characterization Survey at: (website removed)
Appendix D: Pre-lab MLE Activity Worksheets

Pre-Lab Assignments: (1) Crux Vocabulary; (2) Score Sheet for MLE Rubric; and (3) MLE Rubric Notes

Email to: (email removed)

Assigned number  (provide this number to receive credit for your work, place an “N” next to the number if you are not participating in the research):

This document contains all three required pre-lab assignments (for the cohort and case control studies). A brief description of/instructions for each assignment is provided below.

1. Crux vocabulary
   a. Identify and list 4 crux vocabulary words (total) from the cohort and case control studies you read. For each crux vocabulary word list the study it pertains to and provide a brief, specific description how understanding that term or concept can impact the ability to evaluate the study

2. MLE Rubric Score Sheet
   a. Write down the score you are giving to the study based upon the rubric.

3. MLE Rubric Notes
   a. Write notes to capture why you gave the article a specific score in a specific area.

- Email the completed version of this document (all three sections completed) to: (email removed)
- Bring a paper copy of your completed work to lab.
Crux Vocabulary Assignment

All medical research contains vocabulary and concepts that are important and complex, but some terms and concepts may be more important to understand in-depth compared to others. Here you are being introduced to the idea of “crux vocabulary.” Crux vocabulary refers to words or concepts that you must understand in-depth to be able to evaluate the quality of the study.

For example, in the red yeast rice study you read last year, the Food Frequency Questionnaire (FFQ) would be considered “crux vocabulary” because you would need to investigate that term in-depth (beyond the information provided in the journal article) to recognize how that knowledge impacts one’s ability to evaluate the study. Although the survey was “validated,” as mentioned in the article, the original FFQ validation-publication stated that “the multiple R-squared using all factors to predict percent of calories from fat was 0.47” (0.47 is a low correlation). This is important because the FFQ played a large role in eliminating diet as a confounding factor in the study, but the because the FFQ R-squared was so low for determining calories from fat, it raises question about the reliability of the instrument for how it was used in the red yeast rice study. Thus, the conclusion the authors made that changes in diet did not contribute to drops in cholesterol may be called into question.

In this assignment identify and list 4 crux vocabulary words (total) from the cohort and case control studies you read. Use the space below to write your words. For each crux vocabulary word list the study it pertains to and provide a brief, specific description how understanding that term or concept can impact the ability to evaluate the study (similar to the type of description provided above for the FFQ).

Crux Word 1:
(a) Word:
(b) Study(ies) pertain(s):
© how understanding that term or concept can impact the ability to evaluate the study:

Crux Word 2:
(a) Word:
(b) Study(ies) pertain(s):
© how understanding that term or concept can impact the ability to evaluate the study:

Crux Word 3:
(a) Word:
(b) Study(ies) pertain(s):
© how understanding that term or concept can impact the ability to evaluate the study:

Crux Word 4:
(a) Word:
(b) Study(ies) pertain(s):
© how understanding that term or concept can impact the ability to evaluate the study:
MLE Rubric Score Sheet: Write down the score you are giving to the study based upon the rubric.

<table>
<thead>
<tr>
<th>Rubric Field #</th>
<th>Article Section</th>
<th>Description</th>
<th>RCT Score</th>
<th>Cohort Score</th>
<th>Case Control Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>Background</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Objectives</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Introduction</td>
<td>Sub-Total Section 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Study Design</td>
<td>Study Design</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Blinding</td>
<td>Blinding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Write N/A in box if not applicable)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Control/Comparison Group</td>
<td>Control/ Comparison Group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Eligibility Criteria (EC)</td>
<td>Eligibility Criteria (EC)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Generalizability</td>
<td>Generalizability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Participation Rate</td>
<td>Participation Rate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Randomization 1 (From Large Population)</td>
<td>Randomization 1 (From Large Population)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Randomization 2 (treatments)</td>
<td>Randomization 2 (treatments)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Implementation (enrollment, concealment)</td>
<td>Implementation (enrollment, concealment)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Treatment Plan/Intervention 1</td>
<td>Treatment Plan/Intervention 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Treatment Plan/Intervention 2</td>
<td>Treatment Plan/Intervention 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Temporality</td>
<td>Temporality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Measure of Exposure Assessment</td>
<td>Measure of Exposure Assessment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Changes in study Design</td>
<td>Changes in study Design</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Outcomes Described</td>
<td>Outcomes Described</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Measures/Methods Describe/Justified</td>
<td>Measures/Methods Describe/Justified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Quality Enhancement</td>
<td>Quality Enhancement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Sample Size (Power Calculations)</td>
<td>Sample Size (Power Calculations)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Sample Size (Enrolled and Completed)</td>
<td>Sample Size (Enrolled and Completed)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Statistics</td>
<td>Statistical Methods</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Confounding Factors</td>
<td>Confounding Factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Interim Analyses (Write N/A in box if N/A)</td>
<td>Interim Analyses (Write N/A in box if N/A)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sub-Total Possible:
(from left to right)
6/6/6

Sub-Total Possible:
(from left to right)
54/45/36

Adjusted Denominator
Less 3 pts if applicable

Subtract 3 From the Denominator If Wrote N/A

159
<table>
<thead>
<tr>
<th>Rubric Field #</th>
<th>Article Section</th>
<th>Description</th>
<th>RCT Score</th>
<th>Cohort Score</th>
<th>Case Control Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sub-Total (Section 3)</td>
<td>Adjusted Denominator</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 25 | Results/Conclusions (Section 4) | Participant Flow (Write N/A in box if not applicable) | 9/9/9 |              |                    |
| 26 |                           | Key data                                                     |           |              |                    |
| 27 |                           | Numbers Analyzed                                             |           |              |                    |
| 28 |                           | Outcomes and Estimation                                      |           |              |                    |
| 29 |                           | Ancillary Analyses                                           |           |              |                    |
| 30 |                           | Adverse Events/Harms                                          |           |              |                    |
| 31 |                           | Bias/Conflict of Interest                                    |           |              |                    |
| 32 |                           | Interpretation (Consistent with Data)                        |           |              |                    |
| 33 |                           | Overall Evidence/Study Limitations                           |           |              |                    |

| Sub-Total (Section 4) | Adjusted Denominator | Sub-Total Possible: (from left to right) 27/24/24 (less 3 pts if applicable) |

| SCORES | (All 4 Sections) |
| Section 1 | Section 2 | Section 3 | Section 4 |

| TOTAL SCORE |
| Denominator Adjusted | Section 1 | 6 | 6 | 6 |

| TOTAL DENOMINATOR ADJUSTED |

| TOTAL % SCORE | (total score/denominator adjusted x 100%) |

Maximum Scores Possible 96/84/75
Appendix E: MLE Rubric

Medical Literature Evaluation Rubric (RCT, Cohort, Case-Control)

| RANOMIZED CONTROLLED CLINICAL TRIAL (RCT): An RCT is a form of a clinical trial where subjects in a population are randomly assigned to receive or not receive a treatment/exposure/intervention. Both groups are followed-up equally and measured for specific outcomes. | COHORT STUDY: Cohort studies are observational population-based studies that compare a group of people with an exposure to a different group (no exposure or different level of exposure) to answer the question “What are the effects of this exposure?” or “What are the risk factors associated with this exposure?” The studies may be prospective (“where the exposure is defined and subjects are selected before the outcome occurs”) or retrospective (“where the exposure is assessed after the outcome is known, usually by examination of medical records.”). Cohort studies need to observe enough subjects for a long-enough time to generate reliable incident and mortality rates (they are usually large numbers and long-term). Cohort studies are good for identifying incidence and natural history of a disease and “can examine multiple outcomes after a single exposure.” It is the best way to assess absolute risk.

| CASE-CONTROL STUDY (CC): A case-control study is a retrospective observational study comparing cases with a specific disease with a group without the disease (“controls”). These studies investigate an association between the hypothesized exposure and the disease being studied trying to answer the question “What exposures/conditions resulted in the disease?” Case-control studies are designed to estimate odds/relative risk. Ideally controls need to come from the same population as the cases to “reduce the chance that some other difference between the groups is accounting for the difference in the exposure that is under investigation.” Because case-control studies depend upon retrospective data, primarily in the form of people remembering their exposures, recall bias is possible. The use of biologic markers reduce the problem of recall bias.

<table>
<thead>
<tr>
<th>HIERARCHY OF EVIDENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic Reviews</td>
</tr>
<tr>
<td>Critically-Appraised Topics [Evidence Syntheses]</td>
</tr>
<tr>
<td>Critically-Appraised Individual Articles [Article Synopses]</td>
</tr>
<tr>
<td>Randomized Controlled Trials (RCTs)</td>
</tr>
<tr>
<td>Cohort Studies</td>
</tr>
<tr>
<td>Case-Controlled Studies Case Series / Reports</td>
</tr>
<tr>
<td>Background Information / Expert Opinion</td>
</tr>
</tbody>
</table>

**References**


Stefani Dawn, William Troutman, James Nawarskas, Joe Anderson, Stanley Snowden – University of New Mexico, College of Pharmacy 2012
### RCT, Cohort, Case-Control: Introduction Section

<table>
<thead>
<tr>
<th>Description 6</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Background:</strong> Scientific background and explanation of rationale.</td>
<td>Background information is not included.</td>
<td>Background information is included, but some important information is left out.</td>
<td>Most of the relevant or needed background information is included.</td>
<td>Background information is thorough.</td>
<td>1</td>
</tr>
<tr>
<td><strong>NOTE:</strong> It is helpful to look at other similar studies to help determine omissions in background information.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Objectives:</strong> Specific objectives and hypotheses (questions the study was designed to answer).</td>
<td>Does not include objectives or hypotheses.</td>
<td>Includes objectives or hypotheses but are vague, misleading, or are difficult to measure.</td>
<td>Includes measurable objectives or hypothesis but they could be clearer.</td>
<td>Includes clear and measurable objectives or hypotheses.</td>
<td>2</td>
</tr>
</tbody>
</table>

### Introduction Section Score (total points possible 6)

---

6 Descriptions are based upon CONSORT, TREND, SIGN, and STROBE.
### RCT, Cohort, Case-Control: Study Design/Methods Section

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study design:</strong> Provides a details of the study design (e.g. retrospective or prospective cohort, case control, type of trial such as parallel, multi-arm parallel, crossover, cluster, factorial) and conceptual framework (superiority or non-inferiority). <strong>NOTE:</strong> In trying to save space and provide transparency, some journals have authors provide supplementary materials to the publication. When a protocol is provided as supplementary material it is still important that key information about the protocol (in order to judge the protocol) be included in the main body of the publication.</td>
<td>Study design is not described or is missing very basic, important/key information.</td>
<td>Study design description is missing important information.</td>
<td>Study design is mostly described but is missing some information and/or key information is provided only in supplementary materials.</td>
<td>Study design is well described and includes statement of trial type and conceptual framework. Enough detail is provided that it does not require going to supplementary information.</td>
<td>3</td>
</tr>
<tr>
<td><strong>Blinding (masking):</strong> Whether participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If blinding was done, the study describes how blinding occurred and how the success of blinding was evaluated. Blinding is less important for objective measures (e.g., weight) but much more important with subjective measures (e.g., tenderness). Where blinding was not possible, there is some recognition that knowledge of exposure status could have influenced the assessment. Blinding prevents performance and ascertainment bias.</td>
<td>Study does not mention if it is blinded or study is not blinded.</td>
<td>Study claims blinding but does not describe who was blinded, so it is difficult to determine type of blinding. It is probably single blind.</td>
<td>Double-blind study design; study describes who was blinded.</td>
<td>Triple-blind study design; study specifically identifies who was blinded (e.g. group and/or author initials).</td>
<td>4</td>
</tr>
<tr>
<td>Blinding (at least for the patients) is not needed when comparing the pharmacokinetics of two dosage forms.</td>
<td>If blinding was not necessary for the study place and write N/A in the score sheet and adjust the scoring denominator (at the end of the rubric) by subtracting 3 points.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Control/Comparison Group:</strong> Comparison groups need to be selected from comparable source populations in all characteristics except for exposure status. The study needs to provide sufficient details to determine these features.</td>
<td>Study does not adequately describe source population or the source populations are very different between groups (e.g. &gt;50% difference on important confounding factors such as one group having diabetes for a cardiovascular study). Study does not have a control (e.g., comparison group or placebo).</td>
<td>Groups being studied are selected from source populations that are similar but there are differences of potential concern with important confounders (~25-50% difference). Study has a control, but it is not the best choice or there are problems with control design (e.g. administration, placebo taste/ appearance), or control design is not well described.</td>
<td>Groups being studied are selected from source populations that are comparable (e.g. ~25% difference) for most of the important potentially confounding factors. Control for the study appears to be adequate based upon the description, but details are not explicit.</td>
<td>Groups being studied are selected from source populations that are comparable in all respects other than the factor under investigation. Study is very well controlled; or controlled as needed and details are explicit. COHORT: This score is not an option for cohort, as this is a limitation inherent in the study design.</td>
<td>5</td>
</tr>
</tbody>
</table>
### RCT, Cohort & Case-Control: Study Design/Methods Section, cont’d

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligibility criteria (EC) &amp; setting: ECs need to be well described/justified (e.g., method of recruitment, nature and stage of the disease, exclusion criteria). ECs need to be specific like “active hallucinations or suicidal intent”, not broad like “emotional problems” (Humphrey’s et al, 1998). Cases need to be clearly differentiated from controls and the study needs to clearly describe how the cases were selected. Setting and locations where data are collected also need to be well described (e.g., country, city, multi-center, hospital clinic, whether setting might influence results – such as transportation issues). Restrictions should have a sound rationale (e.g. biologic plausibility). If differences between subgroup is unknown, then include relevant subgroups, but plan a priori (Higgins &amp; Green 2008).</td>
<td>Study does not describe EC &amp; setting, or EC are not applied equally to all study groups thereby introducing possible selection bias.</td>
<td>EC are briefly mentioned with little or no justification for the criteria. EC are on the weak side, allowing potential confounding factors. EC might not be applied equally to all study groups. Setting details are not provided.</td>
<td>EC are well described and eliminate the largest and most obvious confounding factors. They appear to be applied equally to all study groups. Justification of EC might be limited. Setting descriptions appear to be complete.</td>
<td>EC are well described, justified, and rigorous, eliminating multiple potential confounding factors. It is clear that the EC have been equally applied to the study groups. Setting is described in detail and carefully considered in the design.</td>
<td>6</td>
</tr>
</tbody>
</table>

- **NOTE:** If a treatment may cause harm to a group, they should be excluded. That aside, there are arguments for and against inclusive and exclusive eligibility criteria. If you exclude too many people does the study population represent reality? Or if the EC are too inclusive do you introduce too many potential confounders? Here are also questions about bias and ECs (e.g., if you let x group into the study can you show a greater effect?) Consider these things, but make sure EC are described, justified (why was an EC applied?) and make sure ECs are equally applied.

| Generalizability: Generalizability refers to how confidently you can apply findings from a study to a larger population. The generalizability of a study may vary by region, thus judgment on this item is based, in part, on your clinical population. Generalizability can impact the generalizability of a study because it determines the nature of the study population. Generalizability is also disease specific. EXAMPLE: A study on heart disease that only has 10% women enrolled is not very generalizable because the proportion of women in the general population that have heart disease is much greater (it is not representative). | Study is very poorly generalizable, it is only for one small subset population. | Study has limited generalizability, there are some key groups missing. | Study is moderately generalizable, covers most of the relevant populations but may have slight weakness in one area. | Study is highly generalizable (sample represents men, women, different ages, ethnicities, relevant to the disease); or it is generalizable to your population of interest. | 7 |

**NOTE:** In general cohort studies tend to have much greater generalizability.
### RCT Only: Study Design/Methods Section, cont’d

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation rate: Differences between the eligible population and the participants are important. Participation rates are calculated by: # participants/eligible subjects. It is best if sorted by cases and controls. It is problematic if there is a large difference in participation rates between the cases and controls, possibly introducing selection bias. If there is little difference between those who accept and decline, then participation rate is less of an issue. Participation rates may vary with type of study. For example surveys typically have low participation rates of around 50%.</td>
<td>The participation rate is low and there is a large difference in participation rates between groups. Is a difference between participants and non-participants, or no data is provided.</td>
<td>Participation rates are moderate and there is a small difference between the rates. There may be a difference between those who participate and those that don’t (but this info may not be provided).</td>
<td>Participation rates are moderately high but nearly equivalent for both groups. Or if there is a low participation rate the authors make an attempt to show similarity between participants and non-participants.</td>
<td>Participations rates are high and essentially equivalent for both groups. Or if there is a somewhat low participation rate the authors conclusively show there is little to no difference between those who accept and those who decline.</td>
<td>8</td>
</tr>
<tr>
<td>Randomization: □ method used □ who did it □ independently performed</td>
<td>Original sample selection is not random, but could have been and study could benefit from it; or sample selection is not stated</td>
<td>Original sample selection is described but is not random and randomization is needed to better control the study.</td>
<td>Original sample is randomly selected from a small population.</td>
<td>Original sample is randomly selected from a large population.</td>
<td>9</td>
</tr>
<tr>
<td>Sequence generation – method used to generate the random allocation sequence (eg, random number table or generator), including details of any special features/restrictions (eg, blocking, stratification), and how code breaking is avoided; Simple randomization is often best but other restrictions may occur in studies with smaller numbers. Allocation by alternation, hospital numbers or date of birth are NOT random. Allocation ratio and concealment: Description of allocation ratio (intended number of participants for each comparison group, eg, 1:1) and method used to implement the random allocation sequence (eg, numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned. Allocation concealment prevents selection bias, third-party assignment is best.</td>
<td>Treatments/group assignments are not randomized, a description of the “randomization” process is not provided, or randomization is claimed but the method is not truly random. The allocation ratio is not described.</td>
<td>Treatments/group assignments are not randomized, and study provides a reasonable and logical explanation. Allocation ratio is provided. If it is an unequal allocation (eg, 1:2) the rationale is not explained or rationale is questionable.</td>
<td>Treatments/group assignments are randomized and methods well described and acceptable relative to participant numbers; Allocation ratio is not well described or justified.</td>
<td>Treatments/group assignments are randomized and allocation ratio is provided; both are well described and justified relative to participant numbers; care was given to avoid bias through inadvertent code-breaking or too many strata.</td>
<td>10</td>
</tr>
<tr>
<td>Implementation: Who enrolled participants, who generated the allocation sequence, who assigned the participants to their groups; and who administers intervention.</td>
<td>The same person(s) enrolled the participants, assigned to treatments, as well as implemented.</td>
<td>The study attempts to separate enrollment, allocation, and implementation; there is overlap in the people who allocate, enroll, and implement.</td>
<td>Implementation of randomization processes are described, some of the same people participated in more than one of the processes (eg, sequence generation and allocation concealment) but not also implementation.</td>
<td>Sequence generation, allocation concealment, and implementation were done by separate people and study reports where investigators stored the allocation list (ideally locked up away from the enrollment location).</td>
<td>11</td>
</tr>
</tbody>
</table>
RCT & Cohort: Study Design/Methods Section, cont’d

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment Plan/Intervention(s):</strong> Provides details of the treatment/intervention to be provided to the study groups including how and when they were actually administered, what was given, delivery method, how subjects were grouped during delivery, who delivered the intervention, setting where the intervention was delivered, and activities to increase compliance/adherence.</td>
<td>Treatment plan/intervention (drug type and dose) is not described.</td>
<td>Treatment plan/intervention is described but contains some weaknesses and/or the issues may affect generalizability (e.g. is a drug no one uses or dose is inappropriate)</td>
<td>Treatment plan/intervention was described and seems to be appropriate, but could have been clearer.</td>
<td>Treatment plan/intervention was well explained, well justified, and logical.</td>
<td>12</td>
</tr>
<tr>
<td>The intervention/treatment follows standard clinical guidelines.</td>
<td>The intervention/ treatment does not follow standard clinical guidelines and reasons are not given to justify the use AND the intervention/treatment does not make sense.</td>
<td>The intervention/ treatment does not follow standard clinical guidelines. It is somewhat justified, but lacks clarity. Treatment may make sense, but raises some question.</td>
<td>The intervention/ treatment follows standard clinical guidelines but its use is not well explained OR if it does not follow standard clinical guidelines it is justified, but lacks some clarity OR if not sure following clinical guidelines, but does appear to be appropriate.</td>
<td>The intervention/ treatment follows standard clinical guidelines OR if it does not follow standard clinical guidelines it is very well justified and makes sense.</td>
<td>13</td>
</tr>
<tr>
<td><strong>Temporality:</strong> Consider the temporality of the study. This is especially important for case-control studies, but applies to all studies. For example, in a clinical trial is there enough time to see the end point? In case-control does the exposure of interest precedes the outcome with the proper time scale (for what is understood about the disease/outcome). If it takes about 10 years for a particular cancer to develop, and the study looks at people 1 year from the exposure, the temporality is problematic.</td>
<td>Temporality of the study does not match the outcome.</td>
<td>Temporality of the study is questionable and little explanation is provided to overcome doubt.</td>
<td>Temporality of the study is reasonably well matched (there is some potential for question) and is explained.</td>
<td>Temporality of the study seems well matched and is justified.</td>
<td>14</td>
</tr>
</tbody>
</table>
### Cohort & Case-Control: Study Design/Methods Section, cont’d

<table>
<thead>
<tr>
<th>Description</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measure of Exposure Assessment:</strong> The study needs to clearly describe how the exposure is assessed and that the measure is reliable. Ideally it should clearly establish that participants have or have not received the exposure and the extent of any exposure. Use of biomarkers (objective assessment) can be a strong prognostic marker (depending on the validity and reliability of the test). Consider the use of subjective versus objective measures for the exposure assessment (see outcomes/end points for objective/subjective descriptions), as well as reliability.</td>
<td>15</td>
</tr>
</tbody>
</table>

- **Exposure assessment is not described and/or uses unreliable methods and/or you cannot reliably discern between exposed/unexposed groups with the methods used.**
- **Prognostic markers or factors for exposure are mentioned briefly but described well enough to judge their quality and/or the study used some markers or factors that have marginal reliability.**
- **Prognostic markers or factors for exposure are fairly well defined and justified. There may be some weak points, but it is unlikely that the study is affected.**
- **Prognostic markers or factors for exposure are clearly defined, well justified and reliable (more objective or training occurs for subjective assessment). One can easily and confidently discern between exposed/unexposed and degrees of exposure.**

| **Changes in the Study Design:** Changes in the study design or unexpected events can occur. These events need to be well described with explanations and/or analyses showing whether or how it might impact results. | 16 |

- **There may have been a change in the study design, but the authors are not explicit about the change and reasons for it. OR The change may introduce significant bias (eg, changing an end-point without clear justification) or have a significant effect on the study’s reliability.**
- **Changes to the initial study protocol or outcomes are described; justifications are included but limited. Changes appear unlikely to introduce bias, but there is some uncertainty.**
- **Changes to the initial study protocol or outcomes are well described, well justified, and unlikely to introduce bias. OR No changes were needed.**
### Outcomes/End Points

Clearly defined primary and secondary end/point outcome measures (including defining objective and subjective measures) and, when applicable, any methods used to enhance the quality of the measurements (eg, multiple observations, training of assessors, use of “gold standards” or “state of the art” measures, whether instruments are validated)

- **a. Examples of objective measures:** assessment of bone mineral density on multiple skeletal sites by DEXA; repeated assessments of pulmonary function.
- **b. Examples of partially-objective:** abstracting ICD-9 codes from administrative database for the diagnosis of osteoporosis; Partially-objective: number of hospitalizations and unscheduled clinic visits from administrative database.
- **c. Examples of subjective measures:** survey to ascertain if a patient has been diagnosed with osteoporosis; asking the patient to report asthma severity (eg., frequency of symptoms, use of rescue medications); tenderness.

**NOTE:** Certain types of studies only have subjective or partially objective measures, this should be taken into consideration and noted below.

Measures need to be well matched to the outcome (reflects what you want), validated, and reliable (this should be described in the study, but may also require additional investigation into the original validation study(ies).) Measures need to be applied equally to all groups.

High quality studies use quality enhancement measures (eg, multiple observations, training of assessors, or use of gold standard measures).

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary and secondary outcomes are not well described or differentiated; there are too many primary outcomes that may contribute to problems with multiplicity of analyses; and/or there is missing a major outcome.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td><strong>Methods used and associated outcome measures are not well described and inappropriate to address objective(s)/ hypotheses.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>The study has no quality enhancement measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Quality enhancement measures are present but limited; they address key areas.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18</td>
</tr>
<tr>
<td><strong>Quality enhancement measures are thorough and rigorous using multiple reliable measures for key end points. Uses gold standard measures.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19</td>
</tr>
</tbody>
</table>
### RCT, Cohort & Case-Control: Study Design/Methods Section, cont’d

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size (power calculations): How sample size was determined. It should include the power/effect size (e.g. 80%, the larger the effect the easier it will be to find), significance level (often alpha=0.05, ideally 2-sided. Alpha indicates probability of a type I error, rejecting null inappropriately), parameters of the event seeking (e.g. 30% reduction in X). Post-hoc calculations of statistical power have little merit.</td>
<td>Sample size justification is not described.</td>
<td>Sample size justifications are described but important details are omitted.</td>
<td>Sample size justifications are described lacks minor details.</td>
<td>Sample size justifications are well described and seem reasonable.</td>
<td>20</td>
</tr>
<tr>
<td>Sample size (enrolled and completed): Study describes the percentage of subjects that dropped out in each group prior to the end of the study. Generally a 20% drop out rate is acceptable, but long-term prospective observational studies may have a higher rate. Consider why people dropped out and whether the dropout rate was similar between groups. Efforts to follow-up with study drop outs to determine reason for leaving and outcomes improves the rigor of the study.</td>
<td>Sample size completing the study falls far below the calculated sample size or if no calculation is provided sample size is very small.</td>
<td>Sample size completing the study falls below the calculated sample size and may impact the study, or no statistical justification/explanation is provided; or no sample size is calculated but it appears the size may be adequate.</td>
<td>Sample size enrolled in and completing the study are very near the calculated sample size. Reasonable statistical justifications are provided for smaller sample size (e.g. effect size is large) and smaller sample size appears to have had little effect on the study.</td>
<td>Sample size enrolled in and completing the study are at or above the calculated sample size.</td>
<td>21</td>
</tr>
</tbody>
</table>

**Study Design/Methods Section Score**

(total points possible: RCT 54; Cohort 45; Case Control 36) ; (3-point denominator deduction possible from this section)
### RCT, Cohort & Case-Control: Results/Conclusions Section

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistical methods</strong>: Statistical methods used to compare groups for outcome(s) are described including methods for additional analyses, such as subgroup analyses and adjusted analyses (eg, for multiple comparisons); distribution of data (eg, normal or skewed) are provided to be able to judge test choices. Confidence intervals (≥95%) and p-values should be included (confidence intervals most importantly); multiple observations require advanced statistics and cannot be treated as independent data; sub-group analyses are clear and use complementary subgroups and a test of interaction (comparison of p-values has high false positives); additional sub-group analyses are discouraged.</td>
<td>No statistics or inappropriate statistical tests/analyses were performed or the study does not name the tests used. No p-values or confidence intervals are provided.</td>
<td>Minimal statistical tests/analyses were performed or minimal descriptions of the tests and why they were chosen is provided. P-values are given, but confidence intervals are not provided.</td>
<td>Adequate descriptions of the distribution of data and statistical tests/analyses are provided; appropriate tests performed. Confidence intervals are provided.</td>
<td>Thorough descriptions of the statistical tests/analyses are provided and justified relative to the distribution of the data and appropriate tests were performed. Confidence intervals and p-values are provided.</td>
<td>22</td>
</tr>
<tr>
<td><strong>Confounding factors</strong>: Primary confounding factors are identified and accounted for in the study design and statistical analysis (e.g. modeling, stratified-, regression-, or sensitivity analysis to correct, control or adjust for confounding factors). Possible presence of confounding factors is a major reason why observational studies are not more highly rated. List in the notes missed confounding factors. An observational study that does not address the possibility of confounding should be given little, if any consideration in clinical decisions.</td>
<td>The study has critical confounding factors that are likely to affect the results and are not addressed statistically or addressed appropriately, if at all by the authors.</td>
<td>Confounding factors are present in the study and may affect the study. Although mentioned, the authors provide minimal explanation.</td>
<td>Provides a list of the major potential confounding factors and takes them into account in the design and analysis, but there may be some concern about potential impact on the study.</td>
<td>Provides a thorough list of potential confounding factors and adequately takes them into account in the design and analysis OR there are few, if any, confounding factors. NOTE: Cohort studies cannot receive a 3.</td>
<td>23</td>
</tr>
<tr>
<td><strong>Interim Analyses (IA)</strong>: If conducted, interim analyses, or examining results as data accumulate, need to be reported (how many there were, what triggered them, stats methods used, stopping rules, if were planned prior to trial start and if not when planned.) Interim analyses may occur to monitor an intervention. They are often done to monitor safety, but can also look at “benefit.” Interim analyses for safety may result in terminating the study early. Interim analyses are often done before recruitment is finished. Timing and frequency need to be specified in the protocol. There are specific group sequential statistical methods are available to adjust for multiple analysis. Data analysis should be done by an independent committee. Performing multiple statistical examinations of accumulating data without correction it can increase the false positive rate.</td>
<td>IA did not occur and was needed OR IA occurred but no information is given about how the data were analyzed; analyses were not pre-specified; analyses were not appropriate and can result in inaccurate results; stopping rules are absent or unclear.</td>
<td>Interim analyses were designed as part of the original protocol, but descriptions or justifications are not well described and it is unclear if appropriate statistical methods were used, stopping rules are well defined.</td>
<td>Interim analyses were designed as part of the original protocol, but descriptions or justifications for interim analyses are less detailed but can infer from them. It appears that appropriate statistical methods and stopping rules are well defined.</td>
<td>Interim analyses were designed as part of the original protocol and use well described, well-justified and appropriate statistical methods. Stopping rules are well defined.</td>
<td>24</td>
</tr>
</tbody>
</table>

No interim analysis occurred/ was not needed - select box & remove 3 points from the denominator.
## RCT, Cohort & Case-Control: Results/Conclusions Section

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Participant flow:</strong> Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned (if applicable), receiving intended treatment (where applicable), completing the study protocol, and analyzed for the primary outcome. Describe protocol deviations from study as planned (including stopping the trial), together with reasons.</td>
<td>Participant flow information is excluded or very poorly described. No explanations are provided for participant dropout and protocol deviations. Diagrams excluded.</td>
<td>Participant flow information minimal. Some brief explanation is provided for participant dropout and protocol deviations, but lacks some clarity. Diagrams excluded.</td>
<td>Some minor details of participant flow are missing, but provides thorough explanation of participant drop-out and protocol deviations. Diagrams are included.</td>
<td>Participant flow is clearly described and justified, especially participant drop-out and protocol deviations. Diagrams are included.</td>
<td>25</td>
</tr>
<tr>
<td><strong>Key Data:</strong> Summarized data is presented clearly and includes: Recruitment: Dates defining the periods of recruitment and follow-up. Baseline data: Baseline demographic and clinical characteristics for each group. Results: Results of all key and relevant variables need to be listed</td>
<td>Key data are omitted in the text and not presented in tables or graphically.</td>
<td>Key data are mentioned either in the text or presented in tables and graphically, but not both and some data details are omitted.</td>
<td>The most important data and some supporting data are discussed in the text and presented in tables and graphically.</td>
<td>All data are thoroughly described and discussed in the text and presented in tables and graphically.</td>
<td>26</td>
</tr>
<tr>
<td><strong>Numbers analyzed:</strong> Number of participants (denominator) in each group included in each analysis. State the results in absolute numbers when possible (e.g. 10/20, not 50%). Sometimes it appears that numbers or data are inconsistent in tables, graphs and/or text (they do not add up) or they are presented in a way that is confusing and it is unclear where certain numbers came from. You do not need to add up all of the numbers in every study (but sometimes it can be helpful), however if you run across such problems in your general reading/review of the study it should be acknowledged.</td>
<td>Did not include the number of participants completing study; Numbers are inconsistent and do not add up (e.g. data in text or tables do not match).</td>
<td>Identified the number of participants in each group; gave results in relative terms (not in absolute numbers); There may be inconsistency or lack of clarity in the numbers/how they are presented.</td>
<td>Identified the number of participants in each group and each analysis. Results are only given in relative terms; Numbers appear to add-up and/or be consistent in their presentation.</td>
<td>Identified the number of participants in each group and each analysis. Results expressed in absolute terms; Numbers appear to add-up and/or be consistent in their presentation.</td>
<td>27</td>
</tr>
</tbody>
</table>
### RCT, Cohort & Case-Control: Results/Conclusions Section

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcomes and estimation:</strong> For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (e.g., 95% confidence interval). Binary outcomes (e.g., death or oxygen dependence) should be presented as both absolute (risk difference) and relative effect (risk ratio/relative risk or odds ratio) sizes.</td>
<td>Results are provided for only the primary outcome and does not provide confidence interval.</td>
<td>Results reported on primary and secondary outcomes and does not provide confidence interval.</td>
<td>A summary of effects is provided for all outcomes; effects size and precision (confidence interval) are provided.</td>
<td>28</td>
<td></td>
</tr>
</tbody>
</table>

**Ancillary analyses:** Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory. One analysis that should be included for RCTs is the Intent to Treat (ITT) analysis which compares patients in the groups to which they were randomly assigned (e.g., based on the initial treat intent), the treatment they actually received, and withdrawal or protocol deviation (thus includes ALL patients, even those who were later excluded or dropped out) (Hollis & Campbell, 1999) If an ITT is not performed it is well justified and describes other analyses (e.g., efficacy or per-protocol analysis) (Greenhalgh, 2006). Case-controls and cohorts should conduct analysis considering those that completed the study and follows up with those that did not complete the study. Multiple analyses of same data create a risk for false positives. Sub-group analyses should be avoided (see statistics section). Pre-specification in trial protocol is more reliable.

| Ancillary analyses are not performed or described. | Ancillary analyses are mentioned but unclear, not well described or may not be appropriate for the study. | Ancillary analyses are well described, but there might be question about the appropriateness of the analysis. | Ancillary analyses, including an ITT in a randomized control trial, are clearly described and well justified. | 29 |

### RCT Only: Results/Conclusions Section

<table>
<thead>
<tr>
<th>Description</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Rubric Field #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adverse events/ harms:</strong> All important adverse events or side effects/ harms in each intervention group.</td>
<td>Adverse events are NOT clearly described.</td>
<td>Adverse events are briefly described and explanations for potential impacts on the study are limited and have some questionable logic, but may slide by.</td>
<td>Adverse events are described, but some details are excluded (e.g., why specific measures were chosen). A logical explanation is provided regarding potential impacts to the study.</td>
<td>Adverse events are clearly described and thoroughly addressed (including how adverse events and side effects were detected and measured, why measured were used, used thorough and appropriate measures, and logical explanations of how the study or conclusions may be impacted).</td>
<td>30</td>
</tr>
</tbody>
</table>

- Consider whether adverse events were detected using passive or active processes.
- Adverse events need to be well defined (What are they specifically looking for and how is it measured?)
- Consider whether adverse events are from treatment or the disease/condition.
- Look for reporting of and consider the rate of adverse events in each group.
# RCT, Cohort & Case-Control: Results/Conclusions Section

| Potential Conflict of Interest: Did the authors declare potential conflicts of interest? Are they affiliated with the company that manufactures the drug of device being tested? | Bias from conflict of interest is clearly present and not disclosed by authors. | Bias or potential conflict of interest is present and disclosed by authors, but it is not disclosed by the authors and is thereby unknown but appears to have little potential influence. | Bias or potential conflict of interest is present and disclosed by authors, but it appears to have minimal impact on study. | Bias or conflict of interest is minimal and is not likely to influence results. | 31 |
| Interpretation: Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision, and the dangers associated with multiplicity of analyses and outcomes. | Discussion/interpretation of the results are inconsistent with data and misleading. | Some of the discussion/interpretations of the results are consistent with the data, but some inconsistencies or overstatements exist. | Most of the discussions/interpretations of the results appear to be consistent with the data. | Discussions/interpretations of the results are very consistent with the data. | 32 |
| Overall evidence: General interpretations of the results in the context of current evidence; Reporting study limitations. | Obvious study limitations are not reported (eg, potential bias and confounding). | Study limitations are mentioned but unclear or some limitations are overlooked. | Study limitations are adequately reported, external validity is discussed, but their potential impact on study results are not well described. | Study limitations are reported, external validity is discussed, potential bias and confounding are acknowledged, and their potential impact on the study results are well described. | 33 |

**Results/Conclusion Section Score (total points possible RCT 27; Cohort & Case Control 24)**

| TOTAL SCORE All sections (RCT 96; Cohort 84; Case Control 75) | TOTAL PERCENTAGE SCORE % Calculated with adjusted denominator |

**REFERENCES FOR RUBRIC DEVELOPMENT:**


Other information obtained from:

Appendix F: Crux Vocabulary

Crux Vocabulary Words Provided by the Experts and Students

Crux Words Identified by the Experts and the “Reason” it is a crux word
Words highlighted in yellow are matches between the expert-identified words and the words identified by the students.

1. **Adjudication** – This concept is very important for the RCT, in this case cardiovascular events were adjudicated by cardiologists, which typically strengthens the confidence in the diagnosis, because there can be quite a bit of variability in diagnosing cardiovascular events. If this is not done then the diagnosis is left up to investigator, which can be a large source of bias.

2. **Adjusted Odds Ratio** – If you just use an odds ratio that is not adjusted for confounders then can get different results.

3. **a-priori power analysis** - They did not do it in this study. The GI events does have an a-priori analysis said needed 143 events only got 55 events because ended study early (did still see some stat diff). No a-priori analysis for cardiovascular. Had 109 events for cardiovascular endpoints, because did not do a-priori - if needed 150 events to see a stat diff – because didn’t have initially can’t tell if there is no interaction. Took Kaplan-Meier said out 6 months but median was only 3 months.

4. **aspirin dose** – The dose range is large at 75-300 mg per day, bleeding risk increases with higher doses, the authors do not differentiate the patients. Use of aspirin and the dose may impact the GI end-points affecting the study results.

5. **bleeding event** - based on their definition it could change the number of patients as an end point. (pg 1911 end points). It is an endpoint-driven study/safety/efficacy. Would expect that Clo + omeprazole would have less bleeding events. Was clearly defined and was clinically reasonable is a bleed would need to intervene on, had some objective measures (2g or 10% hematocrit). Other studies will define moderate/major bleeding event. They are not looking at life threatening.

6. **cardiovascular events** - Appeared to use MACE for defining cardiovascular events but did not state explicitly, how cardiovascular events are defined can greatly impact the interpretations and outcomes of a study.

7. **current use** – (p715, Juurlink study) took people that had an MI – recurrent MI used MACE - initial definition got a prescription of a PPI within 30 days before a MACE event, a few pghs away said they alters definition any prescription of PPI between hospital discharge and a MACE event which could have been up to 1 year. Does not really say when they changed the definition – did not provide rationale of why did 30 days to begin with. Can alter the outcome, makes question validity of the data.

8. **ICD-9 codes** – need to know the weaknesses with using these. Intraoperative variability in coding could have hundreds of different doctors with differing definitions or people entering the codes. If 20% coded wrong could affect the results.
9. **Myocardial infarction (MI)** - Looked at MI not MACE – does not explicitly state how MI defined, used the ICD-10. Good studies criteria fatal versus non-fatal MI and clear definition EKGs blood work. Could come in with unstable angina and could have been called an MI, which affects data interpretation in a large way.

10. **MACE** – used for defining cardiovascular events, MACE is a standardized set of outcomes. It has a mix of harder and softer endpoints. Is important in defining the outcomes for the study.

11. **Percutaneous coronary intervention (PCI)** – bare metal vs drug alluding stent – main difference would cause is duration of treatment (COGENT article, only received follow-up for a few months, for bare metal stenting the minimum duration with Clopidogril is 1 month, if someone received drug alluding stent on clopidogril for a year – followup for COGENT was 3 about 3 month, did they have exclusion criteria for certain stents? All say is coronary stent, guidelines say even with bare metal stent should be on Clopidogril for a year, but - say in article is anticipated taking drug for next 12 months – sometimes PCP if have a bare metal PCP would just take them off of it)
   - primary end point of COGENT study was GI bleeding – if have % of people with bare metal and stop taking Clopid then risk of GI bleeding is going to decrease.

Crux Words Identified by the Students (Ordered from most number of students to least. Highlights indicate a match with the experts.)

1. Kaplan Meier curves (23)
2. Charlson Co-Morbidity Index (22)
3. Cox models /Cox Proportional Hazard Model (18)
4. Conditional linear regression (16)
5. International Classification of Diseases/ICD9-codes (13)
6. cardiovascular death/cardiovascular event/coding for MI/reinfarction (12*)
7. CYP2C19/Cytochrome P450 (9)
8. MACE (7)
9. post hoc analysis (6)
10. stratification/stratified analysis/stratified permuted blocks (6)
11. adherence algorithms/adherent (5)
12. interquartile range (5)
13. medication possession ratio/possession ratio (5)
14. Percutaneous Coronary Intervention (PCI) (5)
15. cardiac risk prediction model/cardiac risk factors (4)
16. pro-drug (4)
17. SAS, version 9.1 (4)
18. nested case control/population-based nested case controlled (4)
19. adjusted odds ratio (3)
20. Concomitant (3)
21. CPT-4 current procedural terminology, fourth edition (3)
22. Canadian classification procedure codes/Classification of Disease and Related Health Problems (2)
23. Clopidogrel (2)
24. Current use (2)
25. FDA current stance on PPI use (2)
26. proton pump inhibitors (2)
27. random sampling with replacement (2)
28. Sampling with replacement (2)
29. IJ50 (1)
30. acetylsalicylic acid (1)
31. angiography (1)
32. attenuating (1)
33. Canadian Institute for Health Information Discharge Abstract Database (1)
34. Claim based analysis (1)
35. Confounders (1)
36. coronary artery bypass graft (1)
37. covariate (1)
38. demographic information (1)
39. forrest plot (1)
40. Helicobacter pylori (1)
41. hazard ratio (adjusted, unadjusted) (1)
42. income quintile (1)
43. index date (1)
44. isoenzyme (1)
45. metabolite (1)
46. misclassifications of exposure (1)
47. multivariable adjustment (1)
48. Omeprazole (1)
49. Ontario acute myocardial infarction mortality prediction rules (1)
50. pharmacy claims (1)
51. polymorphism (1)
52. pooled analysis (1)
53. predicted probability of short term mortality (1)
54. specialty societies (1)
55. standard definition (1)
56. stents (1)
57. universal access (1)
58. validation studies (1)
Appendix G – Scaffold Contributions Survey

1. Please enter your assigned study ID number.

2. Rate the following instructional elements on how well they supported your learning around evaluating different types of medical literature. (no contribution to my learning, a little contribution to my learning, moderate contribution to my learning, high contribution to my learning)
   a. The MLE Rubric by itself
   b. The small group (3-4 people) peer discussion about the RCT
   c. The large class discussion around the MLE Rubric and articles
   d. The crux vocabulary assignment
   e. The class discussion about the crux vocabulary

3. For the pre-lab activities did you work by yourself or with other students? (Don’t worry, there is no right or wrong answer to this – i.e. you won’t get in trouble – it just helps us to better understand how students work outside of class and the contributions of peer conversations plus the MLE Rubric or the MLE Rubric alone). (I worked alone on the pre-lab activities, I worked alone and with student peers, I did all of the pre-lab work together with student peers)

4. Please provide any feedback you may have about any of the different instructional support elements used in this activity (MLE Rubric, crux vocabulary, and in-class discussions). Feedback may include suggestions for changes, what to keep in the activity, or to get rid of to make this the best, most efficient learning experience. Other comments are welcome as well.
References


Chan (Eds.), *Computer Supported Collaborative Learning 2005: The Next 10 Years!* (pp. 331-340). Mahwah, NJ: Lawrence


