The application of epidemiologic research study principles to a variety of research settings

Brandon Keith Mattson
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Veterinary Microbiology and Immunobiology Commons

Recommended Citation
Mattson, Brandon Keith, "The application of epidemiologic research study principles to a variety of research settings" (2009).
Graduate Theses and Dissertations. 10605.
https://lib.dr.iastate.edu/etd/10605
The application of epidemiologic research study principles to a variety of research settings

by

Brandon K. Mattson

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Preventive Medicine

Program of Study Committee
Jared A. Danielson, Co-Major Professor
James P. Toombs, Co-Major Professor
Michael G. Conzemius
Steven Hopkins

Iowa State University
Ames, Iowa
2009

Copyright © Brandon K. Mattson, 2009. All rights reserved.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td>CHAPTER 1. OVERVIEW</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Descriptive Studies</td>
<td>2</td>
</tr>
<tr>
<td>Analytical Studies</td>
<td>3</td>
</tr>
<tr>
<td>Clinical trials and levels of evidence</td>
<td>5</td>
</tr>
<tr>
<td>Examples of epidemiologic research</td>
<td>6</td>
</tr>
<tr>
<td>CHAPTER 2. RETROSPECTIVE COHORT STUDY</td>
<td>9</td>
</tr>
<tr>
<td>Introduction</td>
<td>9</td>
</tr>
<tr>
<td>Methods</td>
<td>13</td>
</tr>
<tr>
<td>Results</td>
<td>14</td>
</tr>
<tr>
<td>Discussion</td>
<td>15</td>
</tr>
<tr>
<td>CHAPTER 3. SYSTEMATIC REVIEW</td>
<td>17</td>
</tr>
<tr>
<td>Introduction</td>
<td>18</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>20</td>
</tr>
<tr>
<td>Results</td>
<td>22</td>
</tr>
<tr>
<td>Discussion</td>
<td>22</td>
</tr>
<tr>
<td>CHAPTER 4. ASSESSMENT MEASURE VALIDATION STUDY</td>
<td>28</td>
</tr>
<tr>
<td>Introduction</td>
<td>28</td>
</tr>
<tr>
<td>Method</td>
<td>31</td>
</tr>
<tr>
<td>Results</td>
<td>36</td>
</tr>
<tr>
<td>Discussion</td>
<td>38</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>42</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>44</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>48</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1: Breakdown of the different study designs for performing research 2

Figure 2: Evidence pyramid showing the level of evidence for study design 6

Figure 3. Proportion of students at each grade level in VCS 456 15

Figure 4. Chi² used to determine if students who take VCS 401 were significantly more likely to get A/A’s in VCS 456 than those that didn’t take VCS 401 15

Figure 5: Breakdown of articles evaluated to determine evidence class of small animal, clinical, orthopedic articles 23
LIST OF TABLES

Table 1. Evidence Classes 21

Table 2. Technical Skills Subscale items with outfit MnSq, logit, 37
and average raw rubric
ABSTRACT

Over the last 3 years the research I was involved in could for the most part be classified in terms of epidemiologic research design. This thesis does not discuss the intricacies of carrying out an epidemiologic study but rather describes the different applications of epidemiologic study designs including the advantages and disadvantages of each. With the ever-growing amount of literature, it is essential to understand the benefits and limitations of each type of study for clinical decision-making and for selecting the appropriate design when attempting to answer research questions. This thesis concludes by giving examples of different studies performed by the author that exemplify specific studies discussed in the introduction.
CHAPTER 1. OVERVIEW

INTRODUCTION

When both performing and reviewing epidemiologic research it is essential to understand various study designs and their strengths and weaknesses. This is particularly important in medicine as decisions are being made for patient care. Ideally these decisions should be based on clinical experience combined with the best clinical evidence provided by systematic research.\(^1\) With the ever expanding volume of literature it is important to not only be able to systematically sift through the research but also to make good choices in scientific methodology to develop valid and applicable results.

There are many important aspects of scientific methodology including: study design, subject selection, study variables, intervention, data collection and analysis, and interpretation of results.\(^2\) Although factors such as time and cost play a role in the choice of study design, the most important influence is the research question under consideration.\(^2\) Since no study design is perfect, it is important to understand the strengths and limitations of each as this may change the answers to the questions and the ability to extrapolate findings.\(^3\)

Figure 1 is a tree diagram showing a breakdown of the different types of epidemiologic study designs. This is not an all inclusive list though it includes the most fundamental designs available which will be the focus of this paper. First, study designs can be broken down into observational and experimental studies. Observational studies are those in which data is being collected without manipulation of the variables within the study. Conversely, experimental studies involve an intervention thereby influencing the variables studied. In the context of medicine, this is performed as a clinical trial. Experimental studies are often the most clinically relevant though ideas for experimental studies are usually derived from observational studies.
An observational study can be further broken down into descriptive and analytical studies. Descriptive studies are designed to determine the distribution of existing variables while answering the who, what, when, where and why without considering a causal relationship.\(^5\) Analytical studies test hypotheses in attempt to determine a causal relationship.\(^2\-^3\) Descriptive studies are most often retrospective (exposure and outcome have occurred by the time of study) in nature while analytical studies may be either retrospective and/or prospective (exposure and/or outcome have not yet occurred by the time of the study).

**Figure 1: Breakdown of the different study designs for performing research**

**DESCRIPTIVE STUDIES**
Case studies and case series are the most basic descriptive studies in that they most often are a report of a new finding whether it is an unusual presentation or a new procedure used to treat a certain disease.² The advantages of case studies and case series is that they are convenient and inexpensive. The disadvantages include a lack of control, and although the case(s) often inspire ideas for analytical studies, no causal inference between treatment and outcome can be made.²,⁶

Cross-sectional studies involve determining both exposure and outcome simultaneously. In essence, you choose a population and then determine whether or not it has been exposed to a certain variable(s) and whether or not they have the outcome under question.⁶ The advantages of cross-sectional studies are similar to case studies in that they are easy to conduct and are inexpensive.² They can also provide useful information such as prevalence or the distribution of variables being studied.² The disadvantages again involve a lack of a causal relationship as you cannot determine a temporal relationship between exposure and outcome, with the exception of exposures that have always existed such as sex or breed.⁵,⁶ Other disadvantages include a possible bias for some studies since only current survivors are studied and it may also under-represent disease of short duration such as those that only occur during a certain season.³

ANALYTICAL STUDIES

The two basic types of analytical studies include case-control studies and cohort studies. In a case-control study you identify a population of individuals with a certain outcome referred to as cases (most often a group of individuals with a certain disease).⁶ A comparison group without that outcome (controls) is chosen and exposure to a certain variable is subsequently determined for cases and controls.⁶ If there is a causal relationship between the exposure and the outcome then there will be a significantly higher (or lower if has a protective effect) amount of
cases exposed than controls. One of the big advantages for case-control studies is that they are
good for studying rare outcomes, such as a rare disease.³ Since the outcome is the starting
variable, you can select all of the people that have already experienced the outcome which
includes subjects over a long period of time. Another advantage of the case-control design is
that it allows for studying many different exposures for one outcome.³ Conversely, case-control
designs do not allow for study of more than one outcome and are not efficient for studying rare
exposures.³ This type of study is also more susceptible to selection bias such that unless the
controls are chosen with the same characteristics (such as socioeconomic characteristics), there
may be confounding variables affecting the results.³,⁶ Finally, case-control studies are also
susceptible to information bias, meaning that since the information is obtained from past records,
the quality and extensiveness of the information may differ from subject to subject.³,⁶

Cohort studies are also considered analytical studies. In contrast to case-control studies
exposure is determined first and the incidence of an outcome(s) of the exposed and non-exposed
second.²-⁴,⁶ Cohort studies can be retrospective in that past information on exposure to some
variable can then be used to look for incidence of an outcome(s) also from past records.⁶ As long
as the exposure is the variable used to determine whether or not there was an outcome then it is a
cohort study. Alternatively, a population of individuals can be followed with some being
exposed to a variable and others not and subsequently determine the prevalence of the outcome
in that population (disease in most instances).⁶ There can also be a combination of the two in that
an exposure can be determined retrospectively and followed up prospectively to determine
incidence of the outcome under question.⁶ The advantage of a retrospective cohort study is that
it is quick, easy to perform, inexpensive and good for studying rare exposures; on the other hand
such studies are susceptible to information bias as there may be differences in quality of past
The advantages of prospective cohort studies are the ability to assess multiple exposures, and the fact that causal relationships can be established more reliably than with other analytical studies because researchers are less likely to obtain incorrect information because the records are being made throughout the study. The disadvantages include biases from non-response or loss too follow-up, higher cost and time commitment, and propensity to diagnose with disease and/or follow the exposed more closely if blinding is not used.

Finally, nested case-control studies are a combination of cohort and a case-control study. A population is chosen from whom baseline data is obtained; the population is then followed for a pre-determined number of years (or until a number of them develop the outcome under question) and a case-control study is then performed.

CLINICAL TRIALS AND LEVELS OF EVIDENCE

As mentioned previously, clinical trials are felt to be the most clinically relevant type of study in epidemiologic research (particularly if they are randomized-controlled trials) in that they are tested on the subject under question. The objective of clinical trials is to test the efficacy of some intervention. Consideration to blinding, randomization, and controls must be taken when performing clinical trials, with the clinical trials that contain all of these features having the highest level of evidence. The disadvantages of clinical trials include their extensive costs, time needed to perform them, and the feasibility and ethicality involved with performing high quality randomized-controlled trials. Again when choosing the evidence from which to base clinical decisions, this does not mean that only randomized-controlled trials can be used. The idea is to use the best available evidence which may only be case studies. On the other hand if there is more than one study to choose from, it is important to understand which studies provide the best evidence on which to base treatment decisions (Figure 2). Systematic reviews have
been shown to have the highest level of evidence while reviews have the lowest level of evidence.

Figure 2: Evidence pyramid showing the studies with increasing levels of evidence.¹⁷

EXAMPLES OF EPIDEMIOLOGIC RESEARCH: CHAPTERS 2-4

In chapter 2 there is an example of a retrospective cohort study. In this study, the population under examination was a group of veterinary students and the exposure was a didactic small animal orthopedic surgery course. Students who had taken the course were compared to students who had not taken the course by the grade received during the 4th year orthopedic surgery rotation. This is a retrospective study in that both the exposure and the outcome had already occurred by the time the study began. I define this as a cohort because the exposure in the population was determined first, followed by the outcome. There is a temporal relationship in that the didactic course came before the clinical rotation. However, we cannot say
conclusively that there is a direct causal relationship due to potential confounding factors discussed in the paper.

In chapter 3 there is an example of a systematic review. Systematic reviews evaluate several research studies on the same topic (most often a specific intervention) with the intent to reveal the best available evidence regarding that topic. This study was a broad systematic review in that it was not evaluating a specific treatment/procedure but evidence value of all small animal orthopedic surgery studies in four separate journals over a five year period. An evidence value was given to all of the clinical small animal orthopedic surgery articles in these four journals and the distribution of the classification was then diagramed (Figure 5). This study indicated that for at least the time period studied (2001-2005) there are very few articles, within the journals studied, with high level evidence.

In chapter 4, the study performed determines the effectiveness of a pre-existing global rating scale at determining technical skills ability. It employs an Item Response Theory model. Its purpose is not to test the efficacy of an intervention, but to test the reliability of an existing rating process against a gold-standard process. This study can be characterized in terms of more than one type of epidemiologic study. First of all, it shares some characteristics with a clinical trial in that two interventions are being implemented with a group of students, and then compared to each other. The interventions in this case would be different assessment methods with one being the “gold standard” and the other the assessment against which we are testing the “gold standard” to determine its effectiveness. In this instance there is a positive control (the “gold standard”) and the outcomes are the scores obtained from each assessment method. Because the purpose of the experiment is to test the reliability of a measure, the same group of students by necessity receives each intervention. This study can also be characterized in terms of
a cohort study in the sense that students have been exposed to 3 years of veterinary school and
the study tests the educational outcome of students’ technical skills. This is unlike a cohort study
in that there is no “un-exposed” group to which we are comparing, because the study’s purpose
is to test the assessment process, and not the educational intervention.
CHAPTER 2. A RETROSPECTIVE COHORT STUDY

THE VALUE OF A DIDACTIC SURGERY COURSE

A paper submitted to *The Journal of Veterinary Medical Education*

Brandon K. Mattson, Richard B. Evans, Jared A. Danielson, Michael G. Conzemius

Abstract

Advanced Small Animal Orthopedic surgery (VCS 401) is a didactic elective offered to third year veterinary medical students. The aim of this study was to determine the value of a purely didactic surgery course. This value was measured by determining how well students who took VCS 401 performed in the subsequent Small Animal Orthopedic Surgery rotation (VCS 456). The course grades of students from VCS 456 were compared among students who participated in, and did not participate in VCS 401. There was a statistically significant difference between the two groups. VCS 401 students performed on average 0.17 higher in overall grade and were 23% more likely to get an A or A- than those without VCS 401.

Introduction

Practitioners expect high levels of proficiency in graduating veterinary students for many different surgical procedures.\textsuperscript{10-12} This expectation may be increasing. Possible explanations for this increase include that a higher frequency of common procedures are being performed or that procedures previously considered uncommon are being performed on a regular basis.\textsuperscript{11} In response to a survey (Greenfield et al. 2004) veterinarians throughout the United States in private exclusively or predominantly small animal practices indicated that the most important skill, procedure, or area of knowledge that new graduates should possess is general and elective surgery. However, Weigel et al. (1992) showed that employers ranked orthopedic surgery and soft tissue surgery as the first and third areas respectively in which
new veterinary graduates of Tennessee College of Veterinary Medicine are most deficient. Similarly, in a more recent study, Walsh et al. (2002) surveyed practitioners to determine whether recent graduates of the University of California-Davis School of Veterinary Medicine were meeting certain expectations. Those surveyed were asked to identify deficiencies from experience with these graduates. The number one deficiency identified by these practitioners (17.3% of total respondents) was “the ability to perform routine surgical procedures, including handling and suturing wounds and lacerations, surgery to remove foreign bodies, spaying and neutering, and administering anesthetics.” Although these last two studies were school specific and may not apply to every school, there is clearly an opportunity for veterinary medical colleges to address these needs in the profession by offering higher quality surgical training to their students.

There are many different methods used to teach required surgical skills to veterinary students prior to their fourth-year surgical rotations. Since the mid 1980’s there has been a trend away from the use of live animals to teach surgery. Some reasons for this trend include: the belief that it is ethically wrong, cost constraints, and faculty time constraints. Despite the method used to teach surgery to students, virtually every method naturally contains some component of didactic teaching. Several studies have compared purely didactic training to differing levels of hands-on training. Chaer et al. (2006) compared catheter-based intervention skills of general surgery residents who received either didactic training alone or didactic training along with training on a computer-based simulator. They found that residents who underwent the simulator training scored significantly higher than residents receiving only didactic training on their first and second catheter-based intervention when graded by surgeons who were
blinded to their training status. Grober et al. (2004) compared junior surgery residents’ microsurgery skills after being trained by didactic training alone versus those that participated in laboratory exercise using silicone tubing or live rat vas deferens. After training, all groups were tested on both silicone tubing and on live rat vas deferens. When assessed by expert, blinded microsurgeons, they found no significant difference in microsurgery adapted global rating scales and checklists on either test between residents who received silicone tubing training and those who received live rat vas deferens training. However, residents who received hands-on training (either silicone tubing or live rat) performed significantly better than those who received didactic training alone on the global rating scales and checklists for both tests. Likewise, they found significantly better anastomotic times among residents who received hands-on training and immediate anastomotic patency rates to be higher with increasing fidelity of training. On the other hand they found no significant difference in anastomotic times, suture placement precision, anastomotic patency rates, and overall quality of the final product when tested on silicone tubing. They also found no significant difference in suture placement precision and overall quality of the final product when tested on live rat vas deferens. Vogelgesang et al. (2002) compared residents’ and fourth-year medical students’ arthrocentesis skills on both a written and practical examination after being trained by either participating in the traditional training method (rheumatology rotation), a didactic lecture, or a program involving a didactic lecture and a hands-on workshop. They found that on the written exam there was no difference in performance between students who received the didactic lecture and those who participated in the program that included both a lecture and a hands-on workshop. However, both of these groups performed significantly better than those who received the traditional training
method on the written exam. They also found that students who participated in the
lecture/hands on program performed significantly better than both comparison groups on the
practical exam, and that the didactic lecture group performed significantly better on the
practical exam than students/residents who went through the rheumatology rotation alone.
This is interesting in that it shows an advantage of a didactic lecture over a rotation in
technical skills training. However it is important to note that individuals in the traditional
training method were only exposed to arthrocentesis when the need arose in a patient
whereas the didactic lecture group was focused solely on arthrocentesis for their training.
Furthermore, the amount of student/resident exposure expected in the traditional training
method was not discussed. Finally, Anastakis et al. (1999) studied first-year residents who
were assigned to a cadaver model, a bench model, or a prepared text for training on six
different surgical procedures after which they were evaluated on their ability to perform each
procedure. They concluded that there was no difference between bench and cadaver training
but both of these groups performed better than the group with only a prepared text.

In veterinary medicine the trend away from using live animals for surgery training has led
to alternative methods such as cadavers, harvested bones, models, spaying/neuter ing
programs, and interactive computer programs. Because lecture settings involve a high
student to faculty ratio, didactic instruction is less expensive to offer than other methods,
particularly hands-on methods. To the researchers’ knowledge, the value of a purely didactic
surgery course has not been studied up to this point.

According to Manu and Lane (1990) there are three successive phases in mastering any
medical procedure. These include: a didactic phase, training phase, and practice phase.
Didactic learning should therefore be considered an important part of surgical training. For
the purpose of this study Advanced Orthopedics (VCS 401), a purely didactic elective, was evaluated.

VCS 401 is a two credit elective course offered to third-year students. The course has a lecture-based format and is co-taught by two board certified surgeons at Iowa State University. The course includes material that covers a wide variety of orthopedic diseases known in veterinary medicine. Although the didactic lecture format is not suitable for acquiring the skills of “complex manual operations,” consistent with Manu and Lane’s suggestion, we would argue that didactic instruction is an important part of the education of veterinary students and we propose that it can have a substantial influence on the success of students in a clinical setting. We hypothesized that students who enrolled in VCS 401 would perform significantly better during the Orthopedic Surgery Rotation (VCS 456) than students who elected not to enroll in VCS 401. Therefore, this study is designed to determine the value of a didactic course in preparing veterinary students for a more hands-on training in the clinical rotation.

**Methods**

Approval was obtained from the Iowa State University Institutional Review Board to conduct this study. The sample included 187 Iowa State veterinary students from the classes of 2004-2006. Students were included in the study if they enrolled in VCS 456. Grades were obtained from course faculty. Student names were converted to an identification number so they could be analyzed without infringing on student confidentiality. Data was analyzed using SAS software version 9.1 (SAS institute inc, Cary, NC). Grades were converted to a number scale with A=4.00, A-=3.67, B+=3.33, B=3.00, B-=2.67, C+=2.33, C=2.00, and C-=1.67. Statistics were reported as average ± standard error. In order to assess
the quality of students who participated in VCS 401, cumulative GPA’s and grades for
Principles of Surgery (VCS 397, required for second year students) were compared among
the two groups for the class of 2006 using t-tests. VCS 456 grades were compared among
students that did and did not enroll in VCS 401 using t-tests and chi-square analysis. Chi-
square analysis was performed to determine if students who enrolled in VCS 401 were more
likely to get A/A-’s than those that did not enroll. Classes were analyzed separately to verify
homogeneity and then combined for the analysis. P-values for statistical tests were reported
with statistical significance set at 0.05.

Results

The class of 2006 cumulative GPA averages for students that did and didn’t enroll in
VCS 401 were 3.22 ± 0.10 and 3.08 ± 0.05 respectively. There was no significant difference
in cumulative GPA’s among groups (p=.25). The average class of 2006 VCS 397 grades for
students that did and didn’t enroll in VCS 401 were 2.97 ± 0.20 and 3.07 ± 0.08
respectively. There was also no significant difference among groups for VCS 397 (p=.58).

Students enrolled in VCS 401 had statistically significant higher average VCS 456 grades
(3.50 ± 0.06) than students not enrolled VCS 401 (3.33 ± 0.04, p=0.01). Figure 1 illustrates
VCS 456 grade proportions for both groups. It shows that students enrolled in VCS 401
received a larger proportion of higher grades than students not enrolled in VCS 401. Figure 2
shows that students enrolled in VCS 401 received a higher percentage of A/A-’s (59%) than
students who did not enrolled in VCS 401 (36%); the chi-square analysis showed this
difference to be significant (p =.0005).
Figure 3: Proportion of students at each grade level in VCS 456. Includes grades of students that enrolled in VCS 401 (blue) and those who did not enroll in VCS 401 (red) prior to VCS 456. No students earned lower than a C-.

Figure 4: Chi$^2$ used to determine if students who take VCS 401 were significantly more likely to get A/A-’s in VCS 456 than those that didn’t take VCS 401. 59% of students that took VCS 401 earned an A/A- in VCS 456 whereas 36% of students that didn’t take VCS 401 earned A/A-. (p=.0004).

Discussion

Although hands-on surgical training is superior to lecture-based didactic training when it comes to practical application, this study shows that there is still value in didactic teaching of principles and skills related to surgery. Despite the differences among groups, because this
study is retrospective in nature, it cannot be determined that the difference was due solely to VCS 401. Other variables might have contributed to the observed difference between groups. For example, students who chose to take the VCS 401 elective might have had more surgical experience than their classmates who did not take the elective. Similarly, students who plan to pursue a career in surgery are more likely than others to take a surgery elective, and are also likely to have a greater desire than their classmates to perform well in a surgery rotation.

Another possible contributing variable may be that students who perform better in the veterinary medicine courses also take the elective, however, because the analysis of cumulative GPA’s and students’ performance in the required VCS 397 surgery course showed no difference between those who did and did not take VCS 401, this explanation seems unlikely.

To determine whether these other variables have an effect on our results, further research is needed to determine whether students who take the VCS 401 elective differ significantly from their classmates in terms of their surgical experience or their desire to pursue a career in surgery.

Conclusion

Students who take the advanced small animal orthopedic surgery elective perform better in the orthopedic rotation than those that don’t. This suggests that didactic surgery courses can be of great value in preparing students for more hands-on surgery experience. This strategy might be particularly useful when there are money/time constraints. Further analysis is necessary to determine if the perceived difference was only due to participation in VCS 401.
CHAPTER 3. A SYSTEMATIC REVIEW

THE EPIDEMIOLOGY OF EVIDENCE: CLASSIFICATION OF PUBLISHED ORTHOPEDIC SURGERY ARTICLES 2001-2005

A paper submitted to The Journal of Veterinary Surgery

B.K. Mattson, R.B. Evans, W.J. Gordon-Evans, M.G. Conzemius

**Objective**- To quantify and describe patterns of research quality in clinical small animal orthopedic surgery.

**Study Design**- Systematic Review

**Methods**- All research articles from JAVMA, AJVR, Veterinary Surgery, and VCOT from 2001 to 2005 were classified as either surgical or non-surgical. The surgery articles were then classified into several categories including: subject species, clinical or bench, and orthopedic or soft tissue surgery. The small animal, clinical, orthopedic surgery articles were then rated on a I-IV scale, I representing the highest level of evidence (systematic reviews) and IV representing the lowest level of evidence (retrospective studies, case reports).

**Results**- Of the 215 small animal, clinical, orthopedic surgery articles 112 were categorized as a class IV, 89 as a class III, 13 as a class II, and 1 article as class I. Conclusions- There was an overwhelming amount of small animal orthopedic articles with low evidence within the journals selected.

**Clinical Relevance**- Although there was a large number of clinical research and prospective studies, the lack of high quality randomized, controlled trials may introduce bias and makes it difficult for clinicians to determine if treatments will be effective.
Keywords

Evidence, randomized, controlled, orthopedic, surgery

Introduction

Though first described in the early 1970’s, evidence-based medicine didn’t take root until the late 1980’s. Since then, medicine based on the “current best evidence” has become a global movement within human medicine in attempt to improve health care. While this may be true for human medicine, evidence-based medicine is still relatively new within the veterinary profession. The necessity of this move toward evidence-based medicine is based on the need to systematically sift through the enormous amount of literature published and the advantage of medicine supported by evidence obtained by scientifically valid methods and study designs.

Although evidence-based medicine is not restricted to randomized, controlled trials (RCT’s), RCT’s, or preferably systematic reviews of RCT’s, provide the best evidence for practitioners to determine the most effective treatment option for their patients. The poor quality and small quantity of RCT’s in surgery has been a criticism of both veterinary and human medicine. Blinded RCT’s lead to unbiased estimates of treatment effects whereas results from studies with inferior study designs can mislead decision making in health care for both individual patients as well as policy-making on a national level. Furthermore, without high quality RCT’s, surgeons are unlikely to change the way they practice. Surgeons are hesitant to change established techniques and procedures and will continue to resist the implementation of new practices when presented with weak data.
There have been several recently published articles highlighting the quality of the published veterinary literature. Brown (2006) investigated how published articles, not limited to surgery, described the process of randomization and the impact of randomization on the quality of the literature. The author concluded that randomization was used in most publications but few of these publications explained their methods of randomization making it difficult to critically review and draw conclusions for patient care. Aragon and Budsberg (2005) published a systematic review of cranial cruciate ligament injury repair in the dog, concluding that there was not sufficient published evidence to favor one surgical procedure over another. Aragon et al. (2007) published a systematic review of treatments for osteoarthritis in dogs concluding that only one drug’s claim to reduce the clinical signs of osteoarthritis has a high level of scientific validity.

The purpose of this study was to quantify and describe patterns of research quality in clinical small animal orthopedic surgery and illuminate successful areas and those that need additional development. To accomplish this task we classified all the articles published in Journal of the American Veterinary Medical Association, American Journal of Veterinary Research, Veterinary Surgery and Veterinary Comparative Orthopaedics and Traumatology (2001-2005) into several categories, analyzed those categories, and further reviewed the small animal orthopedic surgery articles to assess their specific evidence class. As our focus is orthopedic surgery, those four journals were chosen for their impact in the veterinary community or quantity of surgical articles.
Materials and Methods

All articles were cataloged from four journals (named above) for 5 years. These articles were classified using predetermined definitions in a stepwise process (Figure 1). The articles were classified as either surgical or non-surgical. Surgery was defined as “the branch of medicine concerned with the treatment of disease, injury, and deformity by physical operation or manipulation.” Articles that assessed the efficacy of analgesics for surgical procedures were considered non-surgical articles. The surgical articles were separated into orthopedic and general surgery. Orthopedic surgery was defined as “the branch of surgery that embraces the treatment of acute and chronic disorders of the musculoskeletal system, including injuries, disease, dysfunction, and deformities…in the extremities and spine.” Articles that could be classified as both general and orthopedic surgery (i.e. suture studies) were also included in the study. The animal group to which each surgical article belonged (small animal, equine, food animal, comparative, multiple species, exotic) was identified. Each surgical article was then classified as review, bench or clinical research. Clinical research is defined as “Study of drug, biologic or device in human subjects with the intent to discover potential beneficial effects and/or determine its safety and efficacy.” If an article could be classified as both clinical and a bench or clinical and a review, it was included in the study. We interpreted the definition of clinical research liberally, classifying manuscripts as clinical if the research reasonably approximated clinical settings and the outcomes were of direct clinical interest. The articles classified as small animal, clinical, and orthopedic surgery were further reviewed to determine their evidence class using the classification scale provided in Aragon and Budsberg (2005) (Table 1). This is a I-IV scale with I representing the highest level of evidence (systematic reviews), II representing high quality randomized, controlled trials, III representing non-
randomized prospective studies, and IV representing the lowest level of evidence (retrospective studies, case-reports).

Table 1. Evidence Classes. Guidelines for Study Categorization and the Quality of Evidence produced from Veterinary Orthopedic Research (From Aragon CL, Budsberg SC. Applications of Evidence-Based Medicine: Cranial Cruciate Ligament Injury Repair in the Dog. *Vet Surg* 2005;34(2):95. (Reprinted with Permission.)

<table>
<thead>
<tr>
<th>Evidence Class</th>
<th>Study Design</th>
<th>Examples/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Evidence derived from multiple, randomized, blinded, placebo-controlled trials in the target species</td>
<td>Systematic reviews (i.e. meta-analyses). Advantages of meta-analyses are objective appraisal, large number of subjects, improving estimates of association, assimilating large quantities of information, developing findings on a common scale, and improving the quality of primary research</td>
</tr>
<tr>
<td>II</td>
<td>Evidence derived from high quality clinical trials using historical controls</td>
<td>Randomized-controlled clinical studies on animals that developed the disease naturally and is performed in a laboratory setting. Historical controls are thought to be less reliable than randomized controls</td>
</tr>
<tr>
<td>III</td>
<td>Evidence derived from uncontrolled case series</td>
<td>Non-randomized, prospective case comparison studies. Examples include prospective comparison studies that are not truly randomized with limited subjective influence to prospective case series that include subjective clinical impressions to objective gait analysis</td>
</tr>
<tr>
<td>IV</td>
<td>Evidence derived from expert opinion, and/or extrapolated from research or physiologic studies</td>
<td>Retrospective case comparison studies. This category possess the most prominent source of information available today regarding cranial cruciate injuries and surgical repair. Studies on research subjects (non-client owned) are also included in this class</td>
</tr>
</tbody>
</table>
Results

Results are summarized in figure 5 in the form of a tree diagram to show the distribution of articles within each category.

Within the four journals, 3,890 articles were catalogued and classified as either surgery or non surgery. Of those articles, 968 (25%) articles were considered surgical, and 627 (65%) of the surgery articles were classified as orthopedic surgery. Sixty-five percent of the surgical articles were clinical in nature. Of the 627 orthopedic surgery articles, 362 were classified as small animal and 215 of these were further classified as small animal, and clinical research. Fifty-nine percent of the small animal orthopedic articles were clinical in nature. The contribution of JAVMA, AJVR, Veterinary Surgery, and VCOT to these 215 articles was 34, 12, 75 and 94 respectively. Of those 215 small animal/clinical/orthopedic articles 112 were categorized as evidence class IV, 89 as class III, 13 as class II, and one article as class I (Appendix). Forty-eight percent of the small animal/clinical/orthopedic articles were prospective in nature, 17% of which were uncontrolled.

Discussion

High quality RCT’s or systematic reviews of RCT’s are essential to show an unbiased difference in treatment options. This is not to say that all RCT’s are free from bias. Poorly designed RCT’s can be misleading and receive underserved credibility. Consequently, researchers performing meta-analyses need to consider the quality of the RCT’s under study. Brown (2006) found it difficult to critically review and draw conclusions on the majority of the RCT’s since they did not explain their methods of randomization. Despite all of this, RCT’s have been the means by which many useless, dangerous, or deforming surgical procedures have
been discontinued. For example: extracranial-intracranial bypass was a surgical procedure that was described by Yasargil in 1969 for cerebral ischemic disease and stroke prevention. After operating on 100 patients in ten years Zumstein and Yasargil (1980) subsequently concluded that the risk of stroke was reduced from 40% in non-operated cases among current literature to 4.2% in these 100

Figure 5: Breakdown of articles evaluated to determine evidence class of small animal, clinical, orthopedic articles. Based on a I-IV scale with I representing the highest level of evidence (systematic reviews) and IV representing the lowest level of evidence (retrospective, uncontrolled, case-series).
operated cases. Over two decades thousands of patients underwent this procedure with results that appeared to be excellent.\textsuperscript{44} It wasn’t until a large, international RCT was performed did they find that patients that underwent the surgery and received medical management actually had a higher incidence of fatal and nonfatal stroke occurrence than those that only received medical treatment.\textsuperscript{45} Further analysis found that those receiving only medical management also had an improved outcome with respect to several different functional capabilities than those that underwent the surgery.\textsuperscript{46} Due to these results, extracranial-intracranial bypass was virtually eliminated for stroke treatment and management.\textsuperscript{44}

Another example of the effect of a RCT on surgery involved treatment of mammary carcinomas. Less than 20 years ago mastectomy was the gold standard for treatment of primary invasive breast carcinoma.\textsuperscript{47} Numerous RCT’s were performed comparing overall survival and recurrence-free survival between patients undergoing mastectomies compared to patients undergoing breast conservation therapy and found no significant difference between the two procedures.\textsuperscript{47-51} Subsequent long-term follow up of these RCT’s have had similar results.\textsuperscript{52-55}

There are many potential difficulties associated with performing high quality RCT’s in surgery. Some of the more important difficulties that have been reported are:

1. RCT’s tend to be expensive and often there is a lack of funding.\textsuperscript{31,35}
2. Successfully blinding patients, evaluators, and investigators can be difficult.\textsuperscript{31,56}
3. Variability in surgeon experience, skill, and technique making it difficult to standardize procedures and generalize results.\textsuperscript{31-32,35,56}
4. Steep learning curves for certain procedures influences efficacy and timing of randomized controlled trials, particularly when the technique under study is new.\textsuperscript{3-32,35,56}
5. Use of classical control groups is difficult due to the ethical issues arising from sham surgeries.\textsuperscript{31,56}

6. Patient and surgeon preferences make it difficult to randomize.\textsuperscript{35}

7. Rare and urgent/life threatening conditions makes it difficult to recruit, obtain consent, and randomize.\textsuperscript{32}

8. Attaining a large enough sample size to detect a difference that is statistically significant.\textsuperscript{56}

Some potential solutions have been proposed by several researchers. For the lack of funding Shulz et al. (2006) proposed that the American College of Veterinary Surgeons (ACVS) should develop partnerships with health care facilities such as pharmaceutical and pet food companies, as well as implant and equipment manufacturers. They felt that each of these groups has a certain interest in the success of veterinary medicine; pharmaceutical and pet food manufacturers since they market products for pain management and manufacturers of implants and equipment since they provide materials for the surgeries performed. It was also reported that funds could come from diplomats, clients, foundations and industry matching programs.\textsuperscript{29} Another suggestion from Shulz et al. (2006) was to have surgeons demonstrate their commitment by pledging $15 from each cruciate repair. They argue that this would provide a significant amount given that approximately 100,000 cruciate procedures are estimated to be performed annually. Wente et al. (2003) suggested that access to institutional financial resources for those undertaking clinical surgical research should be simplified. Finally, McCulloch et al. (2002) felt that the surgical community should cooperate and create larger groups that could perform certain trials; by so doing they might attract funding.
To help with the difficulty of the learning curve as well as possibly aiding in surgical variability, Went et al. (2003) proposed that all participating surgeons undergo appropriate training before the onset of randomized controlled trials to obtain some standardization. McCulloch et al. (2002) felt that there could be some sort of verification to show competency before beginning randomization. They also proposed using pre-trial success rates and complications as a comparison to test acceptable quality.

Blinding can be very difficult in surgical trials. McCulloch et al. (2002) felt that standardized wound dressings can be successful and also recommended that blinding be used routinely on observers evaluating outcomes.

Despite the difficulty in performing high quality surgical research, several positive aspects were uncovered while assessing the quality of the surgical literature within four major veterinary journals. First of all, a large proportion of articles reported clinical research. Clinical research has an advantage over other types because other research (bench and literature reviews) usually has little immediate impact for improving patient health. Also, over 40% of the orthopedic articles were prospective designs. Prospective studies eliminate biases that may occur in retrospective studies. However, most orthopedic articles were categorized as evidence class III or IV and only 1 article was categorized as evidence class I. This means that the majority of the small animal orthopedic surgery articles in these journals were retrospective and/or uncontrolled. In all retrospective studies it is impossible to determine a causal relationship between treatment and effect. Uncontrolled studies cannot determine if the effect was due to treatment or if it would have occurred despite any intervention. These biases result in low evidence for treatment effects thereby making it difficult for clinicians to determine if the intervention would be effective or significantly better than current methods.
We recognize that the percentages reported in the results section are reflective of the selection of journals, and not estimates of the proportions of all veterinary journals. However, these are 4 major journals impacting veterinarians in the United States.

In conclusion, there is a large amount of small animal orthopedic surgery literature that lacks good quality evidence for clinicians to rely on. This can be improved by increasing the number of high quality randomized, controlled trials and systematic reviews.
CHAPTER 4. ASSESSMENT MEASURE VALIDATION STUDY

A paper to be submitted to *The Journal of Veterinary Medical Education*

Brandon K. Mattson, Jared A. Danielson

**ABSTRACT**

As part of the accreditation process at Iowa State University, development of outcomes assessment of graduating students is in progress. Before implementing major changes in the evaluation of students during their fourth year clinical rotations, assessment of the current evaluation process was performed. One of the nine clinical competencies defined by the AVMA COE, surgical skills, was assessed by comparing the pre-existing grading process, a global rating scale, to scores obtained through blinded reviewing of students videotaped placing skin sutures. Experts reviewing the videotapes showed good inter-rater reliability. The global rating process and the review of videotaped performances produced a similar ranking of students. However, the global rating process tended to overstate student ability when compared with the video observation process.

**INTRODUCTION**

An essential component of every veterinary teaching institution is training students to be proficient in practicing fundamental veterinary medicine autonomously. Paramount in this training is demonstrated competency in performing essential clinical skills. Attributes that both human and veterinary students are expected to have acquired by graduation have been described.  

1-3 The American Veterinary Medical Association (AVMA), through the Council on Education (COE), has defined nine core clinical competencies expected of veterinary graduates as part of the accreditation process of veterinary medical education programs. 4 All colleges seeking accreditation from the AVMA are required to “provide a) the learning objectives for each of the
nine listed competencies, and b) a summary of the analysis of evidence-based data collected for each of the nine listed competencies used to assure that graduates are prepared for entry level practice…

For assessing the nine clinical competencies the AVMA COE mandated the use of outcomes assessment as the basis for program evaluation. Outcomes are defined as consequences of an educational program that can be directly attributed to that program and not outside influences. Outcomes may be short- or long-term, direct or indirect and may also be unintended results though as a rule, intended results are generally measured for assessment purposes. For a number of the institutions publishing their assessment plan, survey instruments have been the most common method by which outcomes are assessed. The surveys obtained responses from alumni, faculty, practitioners (particularly those employing recent graduates) and/or current students with a main focus on assessing the current curriculum’s ability to produce competent veterinarians.

One of the nine core clinical competencies defined by the AVMA COE is “basic surgery skills, experience, and case management.” For the purpose of this study the focus was on performing outcomes assessment of one aspect of this core clinical competency, basic surgery skills. Surgery technical skills have been assessed in various ways. Although little has been published on the methods of assessing technical skills from an outcomes- based approach apart from survey instruments previously mentioned, studies comparing different methods of surgical training have discussed means for assessing surgical skills. In these studies, the use of direct observation and blinded videotape evaluations were used to score study participants. In human medicine, methods used for assessing surgical technical skills have been well defined. The most common methods described include: direct observation, direct observation with criteria,
simulated models with criteria, procedure lists with logbooks, videotapes, motion analysis and virtual reality.\textsuperscript{13-14}

Direct observation without predetermined criteria, as would occur in the operating room during a surgical procedure, is inexpensive and easy to carry out. The drawback of direct observation is that it is unsystematic, unstructured and has a poor test-retest and interobserver reliability.\textsuperscript{13-14} Direct observation with criteria such as checklists and global scores is still inexpensive and though it may be challenging developing specific criteria, it is less subjective than direct observation and can be both valid and reliable.\textsuperscript{13-14} The use of simulated models has been shown to be reliable and valid when coupled with defined criteria.\textsuperscript{13} The disadvantages to this technique is the increased cost and time to obtain and set up. If live animals are used, there are also ethical issues that must be considered. Procedure lists and log books are also very easy and inexpensive to carry out, but lack any indication of level of surgical competency.\textsuperscript{13-14} Alternatively, videotaping specific tasks or procedures with the intent for future evaluation can be expensive and expend a large amount of time but offer constructive feedback and are often highly reliable.\textsuperscript{13-14} Motion analysis and virtual reality also show promise regarding valid measurement of surgical skills.\textsuperscript{13-14}

Other more extensive and less commonly used techniques for assessing surgical technical skills include: objective structured assessment of technical skills (OSATS) which involves a series of stations with simulations designed to assess different techniques and imperial college surgical assessment device (ICSAD) which is a device that used electromagnetic tracking of hand movements to measure manual dexterity.\textsuperscript{13-14} Such devices are currently not readily and affordably available in veterinary medicine.
As part of the accreditation process at Iowa State University, progress has been made at implementing outcomes assessment to evaluate the current curriculum. Prior to making changes to the current grading process of fourth year veterinary students, this study evaluates the ability of the existing process to accurately assess one aspect of the 9 core clinical competencies, surgery technical skills. The current process for assessing surgery skills entails a global rating scale arrived at by consensus of clinicians, residents and technicians. The purpose of this study was to determine if the existing process is a valid measurement of technical skills in surgery.

**METHOD**

**Participants**

The Iowa State University (ISU) Institutional Review Board approved the participation of students in this study. All participants were fourth year veterinary students enrolled in the small animal orthopedic surgery rotation during the 2007-2008 clinical year. Twenty six students participated in the study, including 24 who had been ISUCVM students for several years, and 2 who had transferred for the fourth year from Ross University. Given that the study was performed to evaluate the existing evaluation process as a valid measurement of technical skills and not to evaluate the ISU College of Veterinary Medicine’s (ISUVMC) curriculum, Ross student results were combined with ISU student results for analysis.

**Procedures**

The small animal orthopedic rotation is approximately 2 weeks long and is required for all small animal track students and offered as an elective to the students on the equine/food animal track. Students rotate through the course in groups of 2 to 4. Beginning in August and ending in March all students participating in this rotation were invited to participate in this study. Only one of the students did not provide consent to participate in the study. This student’s reason
for not consenting was due to an aversion for being on a videotape. For consenting students, on the last weekday of the rotation (with an exception of two students who were evaluated approximately mid-rotation) students were videotaped as the skin sutures were being placed at the end of a surgical procedure. Students were often suturing simultaneously in different surgery suites and therefore not every student was videotaped from each rotation. When multiple students were suturing simultaneously, the student to begin suturing first was chosen to be videotaped.

Attention was paid to not videotape any part of the students’ faces, with the bulk of the time focused solely on videotaping the hands. The videotapes were then transferred to DVD for easy viewing. Sound and any identifying information that might link the video segment to a given student were removed in editing. The reviewers were two experienced surgeons who were not involved in orthopedic surgery rotation. Criteria for rating the students on technical skills were determined by a panel of three veterinary surgeons at ISU. Using a 1-5 scale, 1= unsatisfactory/poor, 3= average senior student/satisfactory and 5= as good as or better than an experienced surgeon, students were graded in 6 areas including: efficiency and dexterity using the needle holders, efficiency dexterity using the tissue forceps, accuracy in suture placement, efficiency and dexterity of suture placement, instrument tie technique and efficiency and dexterity of instrument tie technique.

During the rotation, all students were also scored using the course’s pre-existing general 5 point global rating scale (0-4), 4 = outstanding, 3 = good performance, 2 = expected/satisfactory performance, 1 = needs improvement, and 0 = unacceptable level of performance. Toward the end of the rotation, instructors in the rotation met together and, given their experience with each student over the two weeks of the rotation, assigned a score to each
student for each of the following categories: Technical Ability/Clinical Skills, Patient Care, Knowledge/Rounds Participation, Records, and Attitude/Reliability/Initiative. For the analysis comparing the two mechanisms of scoring, the technical ability/clinical skills item was used. Prior to analysis, the 0-4 five-point scale was converted to a 1-5 five-point scale to facilitate the ease of comparison between scales.

Data were analyzed to determine (a) the inter-rater agreement among the two experts who reviewed the video tapes, (b) the relative difficulty of the six sub-items of that scale, and (c) the level of agreement between the video-tape scoring process, and the global rating scale process. All analyses were conducted using a many facets Rasch model with FACETS software.

The multi-facets Rasch Model is represented mathematically as follows:

\[ \ln \left( \frac{P_{nijk}}{P_{nijk-1}} \right) = B_n - D_i - C_j - F_k \]

where

\( P_{nijk} = \) probability of student \( n \) being rated \( K \) on item \( i \) by rater \( j \),

\( P_{nijk-1} = \) probability of student \( n \) being rated \( K-1 \) on item \( i \) by rater \( j \),

\( B_n = \) competence of student \( n \) (in this case, in surgical technical ability),

\( D_i = \) difficulty of item \( i \),

\( C_j = \) severity of rater \( j \),

\( F_k = \) difficulty of rating step \( k \) (on the rating scale) relative to step \( k - 1 \)

Rasch analysis allows for independently analyzing the performance of each rater, item, and examinee, and is well-suited to meaningful analysis of rating-scale data. Unlike many
agreement statistics, Rasch Analysis provides both an estimate of the degree to which raters scored cases similarly (such as is characteristic of joint probability of agreement, and kappa statistics), and the degree to which raters generally moved up or down together (such as is characteristic of correlation-based approaches). Given a specific data set involving raters, examinees, and items, the Rasch approach assumes that raters can be ordered from most to least severe, examinees can be ordered from best to least well prepared, and items can be ordered from most to least difficult. The approach then estimates the probability that any given individual will succeed (as indicated by rater score) on a given item, and seeks to generate the model that most coherently explains all data. Each rater, examinee, and item are assigned two statistics, a fit statistic, which indicates the extent to which that person/item behaved as predicted by the model, and a difficulty (logit) statistic, which provides an indication, relative to other persons or items, of the rater’s severity, the examinee’s ability, or the item’s difficulty.

**Fit**

For this study, Outfit Mean-square (Outfit Mnsq) was calculated as the fit statistic. The Outfit Mnsq is the chi-square statistic divided by its degrees of freedom. The Outfit Mnsq is less affected by sample size than Z-score-based fit statistics, and is therefore a good fit for our relatively small sample size. The expected value for Outfit Mnsq is 1, with values larger than 1 indicating misfit, and values smaller than 1 indicating overfit. Values between 0.5 and 1.5 are considered useful for measurement. Values between 1.5 and 2 are considered unproductive for measurement, but not degrading. Values less than 0.5 are less productive for measurement, but not degrading. Values greater than 2 are degrading.
**Difficulty: The Logit Score**

For the logit score (difficulty statistic) each item, examinee, and respondent are given a score corresponding to its difficulty, competence (or ability) and severity respectively. An average score is 0. Scores are based on success-to-failure odds and the more positive the score, the more difficult/competent/severe and conversely as the scores become more negative they represent easier items, less competent examinees and less severe raters. For example: a respondent of average competence would have a 50% chance of achieving a correct answer on an item that is rated a difficulty of 0. As the logit score (difficulty) for an item increases, the chance that an average respondent would answer correctly decreases, and vice versa. Similarly one would expect respondents with higher competence logit scores to be more likely to answer any particular item (and particularly, more difficult items) correctly than respondents with lower competence logit scores. Likewise raters who systematically assign lower scores than other raters for the same performance are given a higher severity score and vice versa. In Rasch analysis there is no uniformly accepted significant “gap” size between logit scores, however, gaps of .5 or greater are often considered to be meaningful in educational contexts (this difference typically indicating approximately ½ of a grade level).\textsuperscript{73}

**Overall Reliability Statistic**

The multi-facets Rasch Model also calculates an over-all reliability statistic. This statistic is comparable to Cronbach’s alpha and has a minimum value of 0 and a maximum value of 1. A score of .5 indicates that differences between measures are as likely to be due to measurement error as not. A score of .7 indicates a 70% that differences are not due to error, a score of .8 indicates an 80% chance that differences are not due to error, and so forth.\textsuperscript{74}
Analysis

A Rasch analysis was first performed to determine the agreement between the two expert raters in their judgment of the videotaped performances of students. After the results of this analysis provided evidence for good agreement, all six scores from both raters were averaged for each student – providing one “overall” score for each student. Theoretically, this overall score of students’ surgical ability should agree with the global rating scale score for surgical ability used in the existing rotation evaluation process – both providing an indication of over-all surgical technical skill. Therefore, these two sets of scores were compared, again using Rasch Analysis.

RESULTS

Videotaped performance

Raters

The two expert raters who reviewed the videotaped performance both fit the resulting model with Outfit MnSq scores of 1.05 and 0.93. These scores are well within the range considered productive for measurement by conventional standards (0.5 > and < 1.5). The raters also were nearly identical in severity, with Logit scores of -.07 and -.11. These logit scores corresponded to a score of approximately 2.9 on the 5-point rubric.

Items

Table 1 contains the sub-scale scores by item from the expert analysis of the videotapes. None of the Outfit MnSq scores were greater than 2, indicating that none were considered to be degrading to measurement, and all but one (Instrument Tie Technique) were considered productive for measurement (0.5 > and < 1.5). Suture placement/Accuracy was the easiest skill to score well on, with the Instrument Tie Technique being the most difficult skill to score well on. Given the conventional standard of .5 logits being a meaningful difference, most items were
equivalent in difficulty to neighboring items, with the two easiest items being significantly easier than the three most difficult items, and vice versa. The overall rubric score (5 point scale) ranged from 2.74 to 3.10. The overall reliability score for items was .79.

Table 2. Technical Skills Subscale items with outfit MnSq, logit, and average raw rubric scores

<table>
<thead>
<tr>
<th>Subscale Item</th>
<th>Outfit MnSq</th>
<th>Logit</th>
<th>Average raw score (adjusted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suture Placement Accuracy</td>
<td>.92</td>
<td>.45</td>
<td>3.10</td>
</tr>
<tr>
<td>Tissue Forceps Efficiency/Dexterity</td>
<td>.88</td>
<td>.32</td>
<td>3.00</td>
</tr>
<tr>
<td>Needle Holder Efficiency and Dexterity</td>
<td>.78</td>
<td>.08</td>
<td>2.94</td>
</tr>
<tr>
<td>Suture Placement Efficiency and Dexterity</td>
<td>.79</td>
<td>-.21</td>
<td>2.77</td>
</tr>
<tr>
<td>Instrument Tie Efficiency and Dexterity</td>
<td>.85</td>
<td>-.28</td>
<td>2.74</td>
</tr>
<tr>
<td>Instrument Tie Technique</td>
<td>1.69</td>
<td>-.36</td>
<td>2.68</td>
</tr>
</tbody>
</table>

**Examinees**

Of the 26 participants, 24 fit the model, with scores lower than 2. Two participants significantly misfit the model, with scores of 3.4 and 3.1. Ability logit scores ranged significantly, from a high of 2.68 (mean rubric score of 3.82) to a low of –1.46 (mean rubric score of 2.00). The overall reliability score for examinees was .89.

**Video-taped Performance Compared to Global Rating Score (rotation).**

**Raters**

The two rating processes (video-taped and global rating scale) produced a model that provided good fit to the RASCH model, with outfit MnSq scores of .98 and .53. The two sets of
ratings differed considerably in severity. The logit score for the Global Rating Score process was 3.98, and the logit score for the ratings of the Video-taped performances was -3.38 – a difference of over 7 logits. Translated to the rating scale, the average converted global rating score was 4.2 – slightly better than “good performance”. The average videotaped performance rating was 2.8 – slightly poorer than “average senior student/satisfactory”.

**Items**

For this analysis, there was only one item (overall technical skill) so item agreement was not calculated.

**Examinees**

Of the 24 respondents, 20 fit the model with Outfit MnSq scores lower than 2. Four misfit the model with Outfit MnSq scores higher than 2. The overall reliability score was .56.

**DISCUSSION**

With the intent to comply with the AVMA COE accreditation standards assessment of competence of students soon to graduate from ISU is in progress. Before any significant changes were to be made in the grading process of the students during their fourth year clinical rotations, this study was used to determine the ability of the current grading process to assess basic surgery skills, one of the AVMA COE’s defined core competencies. Though direct observation has many drawbacks, as mentioned above the current process involves observation made by a concerted effort between the technician, faculty, and residents over a two week period while the student performs a variety of surgical technical skills.

Videotaping was used as a “gold standard” to which the technical skills score students received during the orthopedic surgery rotation was compared. As the videotapes were taken of students placing skin sutures, this assumes that the ability to place skin sutures is associated with
a student’s overall technical skills. This task was chosen because (a) it is commonly performed on the majority of orthopedic surgery cases at ISU (b) it would be easy to videotape and (c) it is not routinely practiced in the 3rd year surgery course and therefore would have to be learned and performed on surgery rotations during the 4th year. The assumption was that students with superior surgical skills would be able to learn and perform this task better than students with inferior skills.

Assessment of the criteria used to score the skin suturing videotapes revealed good inter-rater reliability. Only one of the six categories was determined to be “unproductive” for measuring, though not “degrading”. Additionally, the overall reliability score was .79 indicating that there was a 79% chance that the difference in measurements was not due to measurement error. Only two students did not fit the model

The fact that both the existing rubric and the video-scoring process fit the resulting RASCH model well suggests that those processes generally rank students equivalently. In other words, students scoring well on the generic rubric also scored well on the suturing task, and vice versa. This validates the current general rubric in the sense that it is a process that is likely to rank students equivalently to the more costly and time-consuming video-tape evaluation process. This did not hold perfectly, however. The fact that four students did not fit the model suggests that there are at least some factors in addition to surgical technical skills that affected rubric scores that were not detected by the skin suturing task.

Additionally, while the two scoring processes produced similar rankings, they produced different overall estimates of ability. On average, students scored a 4.2/5 (somewhat better than good, but not outstanding) on the pre-existing rating scale, while they scored 2.9/5 (slightly worse than an “average” senior student might score) on the video-taped suturing task. The logit
score, which is standardized and based on the relative likelihood of a certain individual receiving a certain score, revealed that for these scales, this was a substantial difference, of nearly 7 logits. There are several potential reasons for this difference. The global rating scale could overestimate skill, perhaps because faculty are evaluating students with lower expectations than those implied by the videotaped scale. Perhaps “outstanding” on the course scoring rubric really means “better than most students”, but does not have any reference to how a professional might conduct surgical tasks. It is also possible that skin suturing is more difficult than an average technical task, and that average skin suturing ability is equivalent to good technical skills overall. In such a case, a skin suturing score may underestimate overall technical skills. Though students are not regularly taught skin suturing, the principals are the same as other suturing methods and students may be expected to have achieved at least an “average” level of competence. Since the videotaping process was meant to involve as little interference as possible, some students were given some aid during the suturing process. The effect of the coaching would likely not change the outcomes of the study. Without coaching it would likely lead to an even more severe scoring on the videotapes widening the severity gap between the two grading scales and supporting an underestimation of technical skills by the global rating scale.

Ultimately, to evaluate competency in students the method of assessment must not only be valid and reliable but also feasible. Videotaping students to allow for evaluation with a pre-determined criteria may be more reliable than direct observation but is not likely feasible for a busy teaching hospital with limited time and money. Conversely, evaluating technical skills through direct observation, though it be through a concerted effort, is often unstructured/unsystematic and ultimately lacks precision and reliability. If direct observation is the sole method by which competency is evaluated, the outcome may not be sufficient to target
strengths and weaknesses in the curriculum. Additional outcomes assessments such as surveying employers of new graduates may be necessary to obtain an accurate account of competency acquired from an educational program.
CONCLUSION

Participation in a variety of different types of epidemiologic studies provides an appreciation for the strengths and challenges of study design in this field of research. With the retrospective cohort study performed it was easy to collect and analyze the data initially. When potential confounders were identified, the study became more difficult as attempts were made to control for those variables. This meant more data collection and analysis and at times the data needed to be collected to control for those confounding variables was not available making inferences from the data less reliable.

In the second study, the importance of good quality study designs and evidence for an overall lack of good quality evidence on which decisions involving small animal orthopedic surgery can be based was apparent (at least for the time period studied). Though there was no research trial conducted, a large amount of time was invested in searching through the literature for studies to include in the systematic review. Furthermore, it can be difficult to determine how many studies are necessary to be representative of the population.

Finally, the third study involved actual experimentation of subjects. As studies involve more people, scheduling and availability become more difficult. Though consent was not a difficult problem in this study, it is easy to see how there may be reluctance in some studies to obtain consent. This makes it difficult to obtain a large enough sample size for meaningful results. On the other hand, while working with the population under question, results seem more meaningful and inferences less overstated.

Through these three research studies the importance of recognizing strengths and limitations of epidemiologic study designs in order to generate quality research is evident. If study designs are chosen and planned appropriately, shortcomings that will effect desired results
can be avoided. Occasionally, due to time or money constraints, these shortcomings cannot be avoided. In these instances recognizing the limitations of these studies will prevent making erroneous assumptions regarding study results.
BIBLIOGRAPHY

9. References
58. Walsh DA, Osburn BI, Schumacher RL. Defining the attributes expected of graduating veterinary medical students, part 2: external evaluation and outcomes assessment.
APPENDIX A

References of article classification for the systematic review

Class 1:


Class 2:

Class 3:


17. Kudnig ST, Fitch RB. Trans-iliais and transsacral brace fixation of sacral fractures and sacro-iliac luxations (seven cases). Halling KB, Lewis DD, Jones RW, et al. Use of circular
external skeletal fixator constructs to stabilize tarsometatarsal arthrodeses in three dogs. *Vet Comp Orthop Traumatol* 17:210-215, 2004


Class 4:


