

Artificial Neural Networks: A systematic review of their efficacy as an innovative resource for
healthcare practice managers.

Jeffrey H. Axt

University of Maryland University College

ProQuest Number:10637978

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10637978

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

TABLE OF CONTENTS

Acknowledgment and Dedication.....	6
Abstract.....	8
I. Chapter 1 – Introduction.....	11
A. Background.....	11
B. The Purpose of this Research.....	12
C. The Management Problem.....	13
D. The Artificial Neural Network.....	13
E. The Value of ANNs to Clinical Practice.....	15
F. Organization and Rationale.....	18
G. Study Design and Approval.....	22
H. Research Questions.....	25
I. Objectives.....	25
II. Chapter 2 – Literature Review.....	27
A. ANN Use in Research.....	27
B. Defining <i>Effectiveness</i> for ANN Applications.....	29
C. Measures of ANN Performance.....	32
D. Prior Work in ANN Application Review in Healthcare.....	36
E. ANN Cost versus Value in Healthcare.....	39
F. The ANN as a Disruptive Innovation.....	43
G. Change Management and New Technologies.....	45
III. Chapter 3 – Conceptual Framework.....	47
A. Managerial Decision Making.....	49

B.	The <i>Nudge</i>	50
C.	The Concept Model Proposition.....	53
D.	Understanding the Potential for ANN Use in Practice.....	59
E.	Dynamic Clinical Decision-making Logic Chain.....	62
IV.	Chapter 4 – Methodology.....	64
A.	Initial Search Protocol Design.....	65
B.	Information Sources and the Boolean Search Criteria.....	66
C.	Abstract Review and Analysis.....	68
D.	Weight of Evidence Analysis.....	70
E.	Weight of Evidence Analysis Questions.....	71
F.	Final WoE Determination.....	74
G.	Data Collection Process.....	74
H.	Data Element Extraction for RQ Analysis.....	76
I.	Examples of Data Extraction.....	78
V.	Chapter 5 – Research Findings and Discussion.....	81
A.	Research Question Data Analysis.....	81
1.	Evaluation regarding research question #1.....	81
2.	Evaluation regarding research question #2.....	85
B.	Unanticipated Data Findings from the Reviewed Studies.....	92
C.	Discussion.....	96
VI.	Chapter 6 –Implications for Clinical Practice and Healthcare Management.....	97
A.	Implications of the Research Question Findings.....	97
1.	Lack of Clinician Familiarity with ANN Tools and Concepts.....	98

2.	Piloting Approach and Practice Development Center	99
3.	Existing Research Demonstrating Practitioner ANN-use Success	101
B.	Post-study Implications for ANNs in Practice.....	102
1.	ANN application delivery	102
2.	Implementing the DRAWN Model.....	103
C.	The Evidence in Support of ANN Value to Practice	106
1.	DRAWN Practice Implications.....	106
D.	Limitations	108
E.	The Healthcare Manager’s Point-of-View	109
1.	Future Research Considerations	110
2.	DRAWN Implications Beyond Healthcare.....	111
F.	The Future of AI Use in Healthcare.....	112
G.	Conclusions and the Future of Healthcare Practice	114
VII.	References.....	116
VIII.	Appendix A – An Overview of ANNs.....	134
A.	Introduction.....	134
B.	ANN Structure	135
C.	Validation and Testing.....	140
D.	Additional Resources	141
E.	The Value of ANNs to Healthcare Research	142
IX.	Appendix B – The PRISMA Statement	145
X.	Appendix C – Predictive Tool Comparison.....	148
XI.	Appendix D – ANN Application Usage in Healthcare Research Publications	149

XII.	Appendix E – Proposition Logic Chain Table	150
XIII.	Appendix F – PRISMA Flow Diagram	153
XIV.	Appendix G – References for Analyzed Studies	154
XV.	Appendix H – Weight of Evidence Analysis Matrix	173
XVI.	Appendix I – Research Question Analysis Matrix	177
XVII.	Appendix J – Coding Examples.....	180
XVIII.	Appendix K – Clinical Applications Tally	220
XIX.	Appendix L – Actual Predictive Power by Study	222

Acknowledgment and Dedication

As with any work of this magnitude there are many whose contributions, criticisms, insights, and sheer patience, that must be recognized. Indeed, this work provides just such an opportunity to bestow such recognition across a broad spectrum of contributors and reviewers. First among them, and with great appreciation for his guidance and tutelage in the dissertation process, would be the input of Dr. Timothy Belliveau, my long-time collaborator and friend who helped to keep the ship on-course despite the seemingly unyielding waves that tossed it about on occasion. His insight and the sometimes-subtle suggestions provided me with the strength to continue when parts of me felt doubt. I look forward to future collaborations with him in furthering this fledgling science of clinical decision making augmented by new, advanced machine-learning technologies. And in that same vein, a great deal of appreciation must be granted to Dr. Paolo Lisboa, who, at a key juncture, gave me invaluable feedback and guidance on where to focus my attention, along with noting some pitfalls to be avoided.

Next, although certainly to no less degree, are my two dissertation advisors, Dr. Walter “Chris” Cason and Dr. John Sherlock. This relationship was in many ways a two-way street, bringing them up-to-speed on the unfamiliar technology of artificial neural networks, while they would stress the necessity for a managerial focus for this research. However, be assured that it was I who benefited from both directions – gaining clarity and refinement to my study while simultaneously preparing me for both the defense and the post-publication arguments that would be required. As we agreed early on, this work has two audiences, those with a more technical lean that might recognize this new opportunity for engaging machine-learning tools, and those from the managerial cadre so well represented at UMUC that were looking for ways to apply research to actual practice (clinical or otherwise). I trust I fulfilled both viewpoints through this

effort, yet I acknowledge that this work barely scratches the surface of what is needed to improve healthcare practice. If it only serves to initiate the conversation in circles that had not yet engaged it, then I consider the effort a resounding success. Yet were it not for these two esteemed doctors and their insightful (and, at times, “incite-full”) commentary, the end product would indeed be of far less value, if of value at all, to its key readers.

And no less gratitude is due to the numerous reviewers and consultants who gave me guidance along the way. Many of those are executives and peers at Hospital for Special Care, my place of employment. To begin, I thank Dr. Roger Thrall, Director of Research, not just for his astute review of my early version of this work, but for his encouragement for its completion. To both Ms. Laurie Whelan, CFO, and Ms. Michelle Milczanowski, Grant Writer, for taking the time to provide their perspectives on financing technologies in healthcare. To Mr. Stanislaw Jankowski, CIO, and the entire Information Technology team, who put up with my sporadic requests for time off to complete major portions of this work. And to Dr. John Votto, CMO and ex-CEO, and others on the medical staff, for their encouragement and feedback on the research as it developed. And, finally, to my editor, Sarah Johnston, who somehow turned my jargon into comprehensible text.

But I would be deeply remiss if I neglected the other side of the table – my family – to whom I dedicate this work. To my wife, Suzanne, who put up with endless absences (for social activities as well as for those mundane chores at home) for four years, continuously encouraging me with words of comfort and confidence over that time. And to my three daughters, Jody, Jessica, and Dawn, for putting up with a “missing father” on those so many occasions when I really should have been there for them. And, finally, to my extended family (Bhakti, Mike, and the gang up in Saratoga Springs, NY) for all their support over those years.

Abstract

Background: Much of the current literature on human decision making in the healthcare setting has documented clear evidence of bias and heuristic thinking in the process of making a diagnostic or prognostic assessment. The evidence suggests that such unrecognized practice errors occur in as much as 15% of all cases (Berner & Graber, 2008). It also suggests that the severity of such errors tends to be high, affecting the patient's long-term outcomes (Graber, Franklin, & Gordon, 2005). A leading topic in the healthcare industry, and in clinical practice management, is around the development and use of machine-learning tools to provide additional guidance to the practicing clinician as a means to improve quality and reduce costs. In addition, while there has been a great deal of research-based use of these tools over the last few decades, there has been almost no study of their usage in direct clinical practice and very little research in their efficacy in actual practice.

Purpose: The purpose of this study is to examine the efficacy of such an innovative machine-learning technology, the Artificial Neural Network (ANN), and to ascertain its value to clinical practice as a diagnostic or prognostic assessment device. ANNs are often employed as classifiers that can determine likely relationships amongst input parameters and generate expected classifications. In a healthcare practice, this would mean providing a clinician with an expected diagnosis or prognosis based on case-specific data given to a trained ANN, and integrating that ANN tool into the clinical workflow. This study attempts to address the gap between ANN use in research versus practice by demonstrating that if the efficacy of the ANN tool is substantiated in the research literature then its application to management practice in the clinical setting is justified. In order to accomplish that, this study was set to determine two aspects of ANN use where they are readily found – in the research literature. A systematic

review of the literature on ANN applications in healthcare was done in order to answer two research questions:

- RQ1: When ANN models have been used in healthcare studies, were they applied *effectively* as a high precision diagnostic or prognostic tool?
- RQ2: Of those ANN studies analyzed, under what conditions and/or what applications have they tended to perform with greater effectiveness (and, conversely, where have they not done so)?

Methodology: A literature search was done for recent healthcare studies that used ANNs as their primary focus, resulting in an initial selection of 364 studies from three separate search groups. After reviewing the candidate selections, eliminating duplicates, and using a Weight of Evidence assessment, the selection set was reduced to 74. Of those, each was examined as to the efficacy and performance of the ANN, as well as its design, context, and clinical application, with the results tabulated to address the two research questions above. The findings of this study indicated that there is strong support for ANN efficacy in research (RQ1), and there is reasonably suggestive value for their application to clinical practice in the area of cancer diagnosis (and even more specifically, in breast cancer diagnosis).

Limitations: The limitations of this study include that there were an insufficient number of studies to offer a stronger response to RQ2. This study also limited its review to only ANN machine-learning tools and not alternative Bayesian applications of that technology. As well, from the review of the extant literature on implementing innovative technologies in healthcare, it was found that financial assessments are exceedingly difficult, especially those technologies which do not lend themselves to a return-on-investment analysis (e.g., their application is not service-based, like an MRI, a diagnostic laboratory test, or a new medication). Finally, this study

did not examine the implementation concerns for ANN technologies, such as resistance to change, although some of those were briefly discussed.

Conclusions/Implications: ANN technology has been used extensively in clinical research studies but has rarely been engaged in healthcare practice. This study provided some measure of support for their use as a clinical consultant for practitioners as a means to reduce error and improve assessment quality. A model of practice implementation for ANNs, the Data Refinery using an ANN with a Nudge (DRAWN), based upon prior work by Gant, Rodway, and Wyatt (2001), was proposed to help guide the clinical manager in integrating ANNs into clinical workflows. This model, treated as a disruptive innovation, could be used in pilot implementation projects as a means of solidifying the value of ANNs to practitioners, which may assist in furthering and expanding ANNs' use in practice. In addition, it was suggested that this model is adaptable to alternative machine-learning technologies, and even to applications outside of the healthcare field itself.

Keywords: artificial intelligence, artificial neural networks, clinical practice, decision-making, healthcare, healthcare management, quality improvement

Chapter 1 – Introduction

Virginia “Ginni” Rometty, President and CEO of IBM, on principles for introducing artificial intelligence technology into healthcare: “The first one is its purpose, to make no mistake that what we are doing is building technology to augment human intelligence, not to replace it, to augment what man does. This is not man versus machine, this is man *and* machine.” (Rometty, 2017).

Background

It is clear from the title that the research done in this paper involves artificial neural networks (ANNs), a topic not common to management study, as will be noted in the literature review. A technology based upon artificially intelligent design, ANNs are machine-learning tools that can assist in decision making. Lee (2017, para. 6) in the NY Times gave a great description for how such AI tools can be applied in any domain of practice:

What is artificial intelligence today? Roughly speaking, it’s technology that takes in huge amounts of information from a specific domain (say, loan repayment histories) and uses it to make a decision in a specific case (whether to give an individual a loan) in the service of a specified goal (maximizing profits for the lender). Think of a spreadsheet on steroids, trained on big data. These tools can outperform human beings at a given task.

Of significance here is that AI tools have rapidly become part of the news and science lexicons, but as with most relatively new technologies, knowledge of its use and its implications have not followed. Thus, as the introductory quote to this paper was meant to illustrate, there is still a great deal of suspicion and concern over those implications. In this study, we examine the use of a particular AI tool in the context of healthcare research in order to ascertain its performance in that research setting, and we do this as a means to justify its use in actual healthcare *practice*,

where such use has been lacking. It is suspected that, using an evidence-based examination (systematic review), and the proposition of an operational model of implementation, clinical healthcare managers might come to see their value to practice in a way not previously recognized.

The Purpose of this Research

The purpose of this dissertation is to examine the application of ANN classification tools, which have been extensively used in healthcare research studies, in order to determine their potential value as a mechanism to assist a care provider in making a diagnostic or prognostic assessment as part of their clinical practice. In this study, the ANN is examined as a tool that can be used to establish correlative relationships between clinical observations (independent variables) and outcomes (dependent variables), replicating a similar effort by the clinical practitioner, and thus providing a consultative guide as a means to improve the clinician's decision making. This paper also proposes a new model of clinical practice that employs the ANN to augment the clinician's decision process. However, that model hinges on the evidenced efficacy of ANN tools as appropriate to practice. Therefore, the primary research focus of this paper is to examine the use of ANN tools to provide classification of diagnostic and prognostic determinants as reported in the research literature and, using a systematic review, to assess ANN's overall success as reported across a decade of studies. This analysis should make an ANN's value clear when used as an innovative technology in a variety of clinical applications, and how can improve the quality of medical care.¹

¹The phrase *clinical application* is used to mean the medical problem or malady for which the healthcare provider would need to make a clinical decision related to diagnosis or prognosis.

The Management Problem

While a detailed review of the extant literature on clinical decision making in practice will be reviewed in Chapters 2 and 3, the literature highlights the problems faced by clinicians. It has been reported that clinical diagnostic decisions are made incorrectly at a rate as high as 15% (Berner & Graber, 2008), resulting in poor or unintended outcomes due in large part to external influences to human decision-making processes (Kahneman, 2011; Pearson & Clair, 1998; Weick, 2005). Given that healthcare represents over 17% of the U.S. economy (Centers for Medicare and Medicaid Services, 2017) and that negative outcomes are costly (Goode, Clancy, Kimball, Meyer, & Eisenberg, 2002) and risk patient lives (Marshall & Milikowski, 2017), that rate of error constitutes a significant burden to clinicians in attempting to improve their practice. The difficulty from a healthcare management perspective is that there are only few objective ways to mediate that problem. Some have suggested greater mental fortitude be applied in the diagnostic and prognostic process (Croskerry, 2002) but with no clear mechanism provided to overcome the influence of human bias and heuristic thinking. Others have suggested the use of checklists and preset guidelines to follow as a means to instill discipline to practice (Atawande, 2009), yet such mechanisms can still play into the cognitive behaviors of even the most skilled clinicians.

Hence, a process or tool that provides the clinician with a higher degree of confidence in their findings seems needed. We begin, however, with a general review of the ANN itself.

The Artificial Neural Network

As employed and evaluated in the present study, an ANN can be defined as a computer-generated software tool which, through a process known as machine-learning, is able to classify some defined set of clinical outcomes based on a set of given clinical parameters; the end-

product of the tool itself is derived through the use of a comprehensive set of historical clinical data. The theoretical concept for ANNs is based on a model called a *perceptron* originally developed by McCulloch and Pitts in 1943 (Lancashire, Lemetre, & Ball, 2009).² The tool's ability to classify outcomes is based on network training with pre-determined gold-standard outcome assessments (which are typically assessed by skilled human clinicians), and it is usually validated post-training by being applied to a different set of data (often from records intentionally held back from being used for training). The post-training validation process is used to ensure that the ANN produces reasonable and accurate outcome relationships to the given parameters. The ANN is conceptually similar to that of a logistic regression tool, but the nature of its processes and algorithms is to establish those relationships in a non-linear manner (that is, the machine-learning algorithms do not attempt to map a correlation like a best-fit line). Like regression tools, ANNs are able to establish correlation but not causality. Unlike regression, the methodology by which an ANN derives the classification is not easily understood; for this reason, it is sometimes referred to as a "black-box" technology (Garson, 1998, p. 16; Lancashire et al., 2009, p. 319). Since the machine-learning approach is not strictly a mathematical formulation (the outcome is *learned* through the data, not calculated through a pre-determined formula), the exact mechanism employed cannot be retraced. Hence, the outcome mechanism remains undeterminable by humans, although it is clearly replicable as established through the validation process mentioned above. (A broader explanation of ANNs is presented in Appendix A.)

While ANN use in research has been well-documented (a literature search covering a ten-year time-frame identified more than 27,000 unique research articles that employed or referenced

² This is discussed in greater detail in Appendix A – An Overview of ANNs.

ANNs), evidence of their application directly to clinical practice is lacking. Hence, in this systematic review across a selected subset of those studies, the present research seeks 1) to determine where ANNs have generally performed well and where they have not, and 2) to ascertain their effectiveness in diagnostic and prognostic applications with the ultimate goal of identifying ANN applications that are particularly well-suited for general clinical practice use. Thus, whereas most previous ANN-based research studies have evaluated only a limited number of clinical applications, the present study will assess them across a broad range of clinical applications, examining that historical evidence through a systematic review.

The Value of ANNs to Clinical Practice

One of this paper's initial reviewers asked, of this research, "Why bother?" That is, what value does an ANN tool bring to clinical practice beyond the simple recognition of correlative factors involved in a diagnostic or prognostic assessment? Simply put, the value may lie in how an machine-learning tool such as an ANN can intercede or intersect with the clinical diagnostic or prognostic decision-making process itself, especially as a means of corroborating or challenging the human decision. Studies that reviewed the clinical decision-making process describe opportunities for such intercession (Croskerry, 2002; Ferreira, Ferreira, Rajgor, Shah, Menezes, & Pietrobon, 2010; Mendel et al., 2011). It has also been suggested that ANN tools, which learn from data rather than being driven by human-derived algorithms, may mitigate the unrecognized bias of human decision makers (Mamede et al, 2010). Thus, it is possible that using an ANN tool as a means to support the clinician may lead to improvements in diagnostic and prognostic decision-making and could result in better patient outcomes, reduced cost, and greater efficiency in the healthcare system overall, which establishes what this study intends to explore.

What is the economic value and cost of ANNs relative to their effectiveness, and does this affect clinical practice? That is, how does the use of ANNs measure up against other clinical decision-making tools, especially in relation to their potential cost, and how might this be measured? As has been noted earlier, ANN toolsets are not often applied in common practice, and they are developed in such a way as to make them not easily applied to standard clinical practice use. Further, while this paper seeks to establish potential economic value, and posits that cost justification could be a significant stimulus for adoption, the evidence in the literature, in either direction, is exceedingly scarce. As a means to resolve this theoretical quandary, I drew on the advice and expertise of several individuals at my current place of employment (Hospital for Special Care, or HSC, a not-for-profit specialty hospital in New Britain, CT) who are intimately involved in the day-to-day operations of that facility, and inquired as to their best practices for examining the financial perspective of new technologies such as ANNs.

In a discussion with Laurie Whelan, Vice President of Finance and Chief Financial Officer at HSC (personal communication, December 12, 2015), Ms. Whelan noted that cost analyses for healthcare technologies differ significantly between non-profit organizations such as HSC and for-profit, commercial institutions,. She explained that Return on Investment (ROI) analysis was atypical in the non-profit healthcare sector, suggesting that it was nearly impossible to determine the actual dollar benefit of any particular technology due to the payment processes used. The relevant payers for most non-profit healthcare institutions (governmental sources such as Medicare and Medicaid) use non-traditional approaches to payment. Medicaid reimbursement programs offer *per diem* inpatient rates which are not related to actual services but reflect a presumptive level of service for which they set the standard. Medicare uses more of a “payment for services” model but with a unique twist; the assignment of a Diagnosis Related Group (DRG)

sets the payment standards for each patient encounter, which can be augmented by payment adjustments for outlier conditions when justified by the institution (and approved by Medicare). Institutions that seek reimbursement are also expected to meet specific quality standards; failure to do so can result in an automatic 2% reduction in payment rates. (Medicare rates are expected increase over the next several years.) The remaining non-Medicare/Medicaid cases are, for the most part, paid through commercial insurance contracts, which Ms. Whelan described as typically case-unique agreements that have as their primary restriction a length of stay limitation, beyond which the agreed reimbursement rate would be significantly reduced (meaning, in effect, that the longer the stay, the less the *per-diem* reimbursement rate, a model which functions as an incentive for reducing length of stay).

These complicated payment models suggest that connecting an investment in technology directly to revenues is extremely complex at best, and typically not viable in public healthcare practice. Thus, the approach most commonly taken to cost control is three-fold:

- 1) Loss/risk mitigation technologies that minimize the risks of hospital care that address seemingly small errors in judgment or care assessment that can result in a disproportionate increase in costs. For example, if a patient develops a pressure wound (commonly referred to as a bed sore) unrelated to the diagnosis, the cost of treatment could rise significantly. Hence any technology that reduces the risk of pressure wounds is potentially beneficial, and would be considered for purchase if not overly costly and if a trial demonstrates value and efficacy.
- 2) Alert-driven systems that warn clinicians of actual or potential risks that could negatively affect a patient's care. Examples include a warning system (pop-up

alert) during the electronic prescription writing process, notifying the clinician of possible drug interaction, or a ventilator device failure alert system that will activate if a mechanical ventilator fails or is disconnected.

- 3) Less interactive is the examination and review of clinical practices to mitigate errors that can extend patient stays beyond their expected periods, such as a situation that can require engagement of additional staff, equipment, or services to accommodate the error.

At the Hospital for Special Care, Ms. Whelan focuses resources on the first two cost control methods, as those tend to have much more direct relationship to process improvement, and the cost savings that result from addressing these can be very significant. Citing an article in the Hartford Business Journal (Pilon, 2015) Ms. Whelan noted that Connecticut has one of the highest premiums for malpractice insurance in the nation, suggesting that the cost of insurance increases the overall cost of risk, and that technologies that reduce risk can offset high insurance costs. She also recognized grant money as an additional, though uncommon, source of technology funding.

Accordingly, I asked HSC's primary research grant writer, Michelle Milczanowski (personal communication, December 8, 2015), about the conditions funders typically require to justify requests for healthcare technology funding. Ms. Milczanowski indicated that there were really no defined standards of practice, with many grant funders asking for evidence of improved quality of care or a better patient outcome as measures of value, rather than some monetary indicator. She observed that, in her years of experience, she had only rarely encountered grantors requiring evidence of financial return as part of funding agreements, and those cases were almost exclusively oriented toward for-profit organizations such as pharmaceutical firms or

large corporate centers of care (e.g., wellness or urgent-care groups). These observations aligned with Ms. Whelan’s comments about the challenges of quantifying financial benefits of healthcare technology innovation within a non-profit organization.

What remains, however, are the three cost control approaches outlined by Ms. Whelan, and in the present study, these will be useful in assessing the value of the ANN tool in the nonprofit healthcare environment. Given that the nonprofit healthcare sector has widely adopted risk reduction and error mitigation as acceptable means for improving outcomes and reducing costs, if it can be demonstrated that ANNs can be valuable in risk reduction and/or error mitigation, then the implementation of ANNs in clinical practice can be justified. A number of studies have, in fact, suggested that the inclusion of clinical decision support systems into practice would provide opportunities for risk control and error mitigation (Croskerry, 2002; Ferreira et al., 2010; Jaspers, Smeulers, Vermeulen, & Peute, 2011; Smith, Saunders, Stuckhardt, & McGinnis, 2012). This topic will be explored further in Chapter 6 (Implications for Clinical Practice).

Organization and Rationale

As part of this systematic review, an explanation of how a proposed theoretical model for generalized ANN usage will be given – a model that will be examined in how it may attain both clinical and financial value from its adoption. The steps to achieve that argument are to explicate the proposed model and how this study informs it, to provide foundation to that model through a systematic review which demonstrates ANN effectiveness, and to examine the implications to practice when such a model is employed. Thus, this paper is organized in the following manner:

- Chapter 1: Introduction – An overview of recent healthcare research that used ANNs in healthcare and suggested their potential value to clinical practice.

- Chapter 2: Literature Review – An examination of published peer-reviewed studies about ANNs and their usage, including a review of any similar systematic reviews in the literature, followed by discussion of how this paper fills the gaps left by those studies. Also examined in this chapter is the literature pertaining to managerial aspects of employing new, innovative, and even disruptive, technology, exploring the value, use, and cost of ANNs within those contexts. Finally, we review the literature pertaining to change management and how that might play into ANN adoption and use.
- Chapter 3: Conceptual Framework – An explication of the conceptual model that defines the basis for the proposition of this study, and how that model might potentially be integrated into clinical practice. The literature relating to human decision making is explored and the way in which those extant theories affect the clinical decision process is examined. Finally, the true potential of ANNs in clinical practice, given the theoretical models discussed above, is explicated to provide some sense of the value-add potential that they might bring to that practice.
- Chapter 4: Methodology – An examination of the review tools used by this paper, including the literature search protocol, the criteria for selection of studies to be examined, and a summary of how the selected studies will be examined.
- Chapter 5: Research Findings – An explanation of the findings from the systematic review, including examination of those findings in relation to the research questions posed at the outset. This also includes using some statistical analytics, to give some assessment of how those findings inform us about the strengths and weaknesses of those findings as they pertain to the research questions.

- Chapter 6: Implications for Clinical Practice – An examination of how the findings described in Chapter 5 could directly affect current clinical practice, with specific attention to the conceptual model explained in Chapter 3, and recommendations for how these findings might be used to improve clinical decision-making, with particular focus on what a clinical manager would need to do in order to employ them. Included here as well is the identification of limitations of this study, and what additional opportunities for research have come from the work done here.

As noted earlier, a preliminary review of the literature failed to provide evidence for the general use of ANN tools in clinical practice (Gant, Rodway, & Wyatt, 2001, p. 329; Lisboa & Taktak, 2006, p. 411). In determining ANN effectiveness, an early review (Lisboa, 2002) suggested that ANNs had been used successfully in studies involving three application domains (oncology, critical care, and cardiovascular medicine), but that review was limited in scope and was thus deemed insufficient to address effectiveness in a greater breadth of practice application. Thus, taken together, the need for this paper is established – that is, to provide information and recommendations to assist clinical practice managers in employing ANN tools across a broad spectrum of practice domains.

As computer technology has advanced, so has ANN application development, as reflected in the increased availability of software products that support them. Indeed, many widely-used statistical software programs (e.g., MatLab, SAS, and SPSS) support ANN toolsets in their current releases, making such technology more accessible and more likely to be used. As Reio (2009) suggested, emergent research methods arise when more traditional methods are inadequate to the research question. The growing availability of ANN toolsets represents an

opportunity for just such an emergent technology, "...creating the need for new ways of thinking...on the part of researchers facing organizational problems" (Reio, 2009, p. 144).

For example, the "new way of thinking" might include examining ways to use ANNs within clinical practice, evaluating their integration into the clinical decision-making process itself. Since most existing ANN research was undertaken to demonstrate efficacy or value for an ANN tool under specific conditions (e.g., in a particular diagnostic assessment), what is now needed is to develop a model defining the use of that tool under a broader range of clinical practices, and to do so in a flexible way that supports integration into the clinician's general workflow. The requirement for flexibility challenges the traditional research approach, however; those challenges are examined in some detail in Chapter 2 and in the associated introductory review of ANNs in Appendix A. Thus, if ANNs are to be directly applied to clinical practice, then their actual value must be clearly established through a trail of evidence, which defines the primary research task for this paper.

Study Design and Approval

This study and its methodology (a systematic review) were proposed to The Graduate School at the University of Maryland University College (UMUC) Department of Management, and the proposal was approved for study. A research review team was assembled to assess this study's design and to ensure compliance with quality and transparency measures in order to meet the University's guidelines, a process suggested by several authors (Gough, Oliver, & Thomas, 2012, p. 8; Oliver, Dickson, & Newman, 2012, pp. 79-80). To conform to the UMUC requirement for alignment of the dissertation topic with the Academy of Management's (2007) division and interest group domains, two topic divisions were selected:

1. Healthcare Management

- a. Relates to “...performance of health care workers and organizations; public policy issues, such as...quality of care, and their implications for managing health care organizations” (para. 7).
 - b. In highlighting the effectiveness of ANN tools used in research, it is expected that healthcare performance measures for treatment and assessment will be improved, with the impact being a reduction in time needed to recognize a disease process, as well as improved accuracy (diagnostics) and in achieving greater accuracy in estimates of therapeutic outcomes (prognostics).
2. Technology & Innovation Management
- a. This domain as it relates to “process technologies” and includes, “...innovation diffusion and the development, implementation and use of technologies...[and] organizational processes by which technically-oriented activities are integrated into organizations” (para. 25).
 - b. This study serves to recognize the value provided by using a new technology (ANNs) within the healthcare domain and across a variety of disciplines (disease and injury processes) in order to gain improved predictions of outcome.

In order to comply with the UMUC requirements for a systematic review, a PRISMA Statement (Liberati et al., 2009) is used as a design model for this study. The PRISMA Statement is founded on a 27-item checklist and a four-phase flow diagram, and both highlight those items that were deemed essential for transparency (Liberati et al., 2009, p. W-65). According to its designers, the PRISMA Statement was developed to ensure transparency and complete reporting of systematic reviews, and to “...help authors report a wide array of systematic reviews to assess the benefits and harms” of interventions (Liberati et al., 2009, p. W-

66). The ANN tool is not, itself, an intervention, but it can aid in assessing what interventions are clinically appropriate to the patient encounter (diagnostics) and to set expectations for clinical therapeutics (prognostics), making the PRISMA Statement process a well-founded and suitable model upon which to base this review. For reference, a PRISMA 2009 Checklist is provided as Appendix B, and is annotated with references within this dissertation. Some minor adjustments to the PRISMA process were warranted, given the current study's focus on diagnostic/prognostic assessment, but as Liberati et al. recognized, "...authors who address questions relating to etiology, diagnosis, or prognosis,...and who review epidemiological or diagnostic accuracy studies may need to modify or incorporate additional items" (2009, pp. W-66-W-68). Any such adjustments are clearly identified within the paper as well as being noted on the checklist in Appendix B.

Next is an examination of the value of using a systematic review process, and why that approach (as opposed to other statistical or analytic models) has been implemented as part of this paper's methodology. First, aside from the requirements prescribed by the PRISMA Statement (Liberati et al., 2009), the overall structure of the systematic review has been well documented and researched. While there are minor design differences between systematic review models (Liberati et al., 2009, p. 681), Petticrew and Roberts (2006, p. 27) provided a standard set of stages required, hence that set is also employed as a general guide here. The intent of this study aligns with the aim of the systematic review process, that being, "...to locate, select, and appraise as much as possible of the research relevant to the particular review question(s)" (Denyer & Tranfield, 2009, p. 683). As Petticrew and Roberts (2006, p. 35) suggested, when a field is immature (as healthcare ANN application is), a systematic review can highlight the absence of empirical underpinnings. Further, systematic reviews can identify gaps and aid in

directing future research studies (Petticrew & Roberts, 2006, p. 25), benefits that are expected to be derived from the present work.

Research Questions

As any systematic review requires, the process starts with the determination of the research questions to be answered (Briner & Denyer, 2012, p. 117; Denyer & Tranfield, 2009, p. 681; Petticrew & Roberts, 2006, p. 27; Rousseau, Manning, & Denyer, 2008, p. 501). As suggested by Rousseau et al. (2008, p. 501), the process begins with the end in mind; hence, the review question(s) must reflect the review's intended purpose. Thus, the following two research questions are posed:

- When ANN models have been used in healthcare studies, were they applied *effectively* as high-precision diagnostic and/or prognostic tools?
- Of the ANN studies analyzed, under what conditions and/or using what applications have they tended to perform with greater effectiveness, and, conversely, where have they not been as effective?

Objectives

Earlier ANN evaluative studies were reviewed to inform this exploration of ANN-based research and to explore the limitations and challenges of studies using ANN technologies. The intended outcome of the present study is to achieve what Bethel and Bernard (2010, pp. 232-233) called the goal of research synthesis and what Harden and Thomas (2005, p. 260) identified as the most common reason for a systematic review – that is, to answer the “what works?” question by aggregating and analyzing existing research evidence. In this particular case, “what works?” is determined by how generally effective ANN-based tools are reported to be (Research Question #1) and to identify where they performed better or worse (Research Question #2). These

answers are expected to provide a rationale for expanding, or not, the use of ANN technology in clinical practice application, and to identify what capabilities, if any, ANNs can bring to bear on clinical decision-making. In addition, if this information is incorporated into the proposed concept model, the resulting product could function as a guide for clinical managers and systems developers in the deployment of clinical ANN applications. Finally, the outcomes of this analysis will be tallied and synthesized to establish not only whether an ANN toolset is viable, but also whether it is productive (in terms of operational efficiencies). The final objective is to determine if ANNs can be applied directly to healthcare practice as an assistive assessment tool that could considerably influence the delivery of patient care, and yet not replace clinicians nor their accumulated experience and skill. If ANN diagnostic and prognostic clinical assessment tools are made available using the model presented, they might give healthcare providers, who often have an array of advanced technologies around them, a system design for incorporating ANN technology into their practice. The model, when integrated into healthcare medical record applications, can also be employed by those who have little more than mobile devices with general Internet access, representing a quantum leap in the distribution of assistive clinical knowledge. However, these concepts can only be explored when the research has completed (that is, in Chapter 6).

As noted earlier, a general description and review of ANN technology is provided as Appendix A. This is not a technical treatise on ANN technology; rather, it provides general information on how ANNs function; summarizes their advantages and disadvantages as compared to other analytical and statistical toolsets; and suggests how they can be applied to clinical decision support and problem solving in a manner that would elicit confidence and understanding by practicing clinicians.

Chapter 2 – Literature Review

In this chapter, we review the literature to identify evidence of the effectiveness of ANN applications in research studies and to provide a baseline for the development of the conceptual framework used in this study. We also examine the implications of ANNs from a new technology perspective – what does the literature indicate regarding the use and implementation of new technologies.

ANN Use in Research

In clinical research studies, ANN tools are commonly used as classification devices, the second of three healthcare use models identified by Lisboa (2002, p. 13). Using that approach, given a set of patient-based parameters (values), an ANN tool can identify a class or grouping into which those values would place a particular patient's case. For example, one might choose parameters (the input variables) for temperature, headache pain, sore throat pain, and post-nasal drip, and associate those with a Yes/No determination for an influenza diagnosis (the output variable, or the classification). Together, the values that consistently map to an output of "Yes" would be the classifier for influenza, hence making that ANN a diagnostic tool. This paper examines ANN utilization within research studies that use only this type of application, doing so through a systematic review process.

However, there are limitations to this type of ANN usage. As Garson (1998, p. 16) warned, neural networks lend themselves to prediction but not to causal analysis, relegating their use to correlations between the input and output variables, much as logistic regression in more traditional statistics. Garson (1998, p. 16) also recognized, as we noted in Appendix A, that, at best, it can be very difficult to understand how neural nets arrive at their results. Lisboa and Taktak (2006, p. 408) expressed this point as creating a concern related to transparency,

necessitating that the researcher explain "...what influences the network predictions and how to resolve outcome predictions in terms of readily understood clinical statements." Gant et al. (2001, p. 350) suggested that the actual methods of ANNs can never truly be analyzed, and that the ANN network building process is so detailed that it cannot be understood in a direct way by a human observer. Gant et al. (2001, p. 350) further stated that it is important for ANNs to be applied in circumstances where their accuracy can be assessed, such as where random quality control checks can be performed to validate them, likening the process to that used in randomized controlled trials (RCTs). However, not all studies lend themselves to RCT analysis, as has been recognized within the Cochrane Collaboration's Handbook for Systematic Reviews of Interventions (Reeves, Deeks, Higgins, & Wells, 2011). Indeed, Gant et al. (2001) attempted to address this more globally by providing a comparative analysis between three predictive tool methodologies: ANNs, statistical models, and knowledge-based (expert) systems. That table (reproduced here as Appendix C) illustrates the strengths and weaknesses of ANNs in relation to other methods of data classification, which Gant et al. (2001) judged using five criteria: accuracy, generality, clinical credibility, ease of development, and clinical effectiveness (p. 351). It is noteworthy that according to Gant et al.'s (2001) analysis, ANNs completely fail the "Clinical credibility" criterion, while the other methodologies achieve something much closer to a passing grade.

Thus, researchers and practitioners must overcome several significant challenges before ANNs can be accepted as a predictive and classification tool, at least within the healthcare arena: the tendency of ANNs to predict rather than analyze; a lack of understanding of how ANNs arrive at their results; the need for a quality control process; and the combined effect of these factors on clinical outcomes. Indeed, Gant et al. (2001) expressed these concerns directly:

We believe that the very plasticity of ANNs as regards not only their internal architecture but also their adaptability to different data sets is also their Achilles' heel. It is exactly these extraordinarily wide ranging novel potential applications that bring with them equally novel and complex considerations of not only where they fit in the clinical decision-making algorithm, but also how ethically and legally acceptable their implementation might be. (p. 331)

What could mitigate these challenges? How can the major limitations of ANNs be addressed in order to promote a more generalized acceptance in clinical practice? As suggested by the authors above, the answer lies both with the size of the dataset upon which the neural network is trained as well as the methodology used to train the ANN, both of which are addressed in the next section of this chapter.

First, if one could demonstrate with a reasonable degree of reliability that ANNs have been successfully employed in a variety of clinical applications, then it would be justifiable to claim that such broad spectrum success of learning algorithms and the parameters they employ might not be particular to local factors, but instead relate to some intrinsic aspect of ANNs as a whole. It may be possible that the successful utilization of ANNs within research speaks to their value, even if one cannot define a precise mechanism to illustrate the ANN learning process. Hence, such evidence could suggest that the value of ANNs can be demonstrated through replication of ANN research studies. Thus, we expect to examine in this dissertation what the evidenced effective use of ANNs within research shows, which might also address the concerns over the reliability of ANN performance in practical applications.

Defining *Effectiveness* for ANN Applications.

This analysis requires some method of assessing the *effectiveness* of ANN applications in the studies under examination. While the need for a basic quality assessment of each study is clearly required (for example, using a Weight of Evidence assessment), three guidelines have been recommended to evaluate ANN effectiveness (Adya & Collopy, 1998; Collopy, Adya, & Armstrong, 1994):

- 1) Compare to accepted models: Compare the resulting ANN outcome with well-accepted models, noting that the ANN should perform at least as well. For example, if the ANN notes a strong relationship between one or more inputs and the output variable, a standard regression formula could be used as a means of validating that relationship.
- 2) Validate the samples: Out-of-sample validation is essential, since validation using a non-randomly selected subset of the sample dataset is, at best, suspect. There is no way to determine if the cross-validation subset mirrors the training set too closely, resulting in the samples used as unrepresentative of the full population, which could potentially lead to overfitting of the training set (as discussed in the Appendix A section on validation and testing).
- 3) Ensure adequate sample size: The sample size must be sufficient to demonstrate reasonable representation of the full population being studied, and any constraining factors used must not be relevant to the variables being evaluated. Conversely, attempting to validate the sample set by using representative values that are not pertinent to the study would also be spurious (for example, using Age and Gender distribution to validate sample selection would be inappropriate if the outcome

measure is unaffected by either of those variables). These are often referred to as confounding issues (Donaldson, 2012, p. 256; Kukuli & Ganguli, 2012, p. 2012).

For the purposes of this study, we will assume that the first factor (comparison to accepted models) has been assessed by the author(s) of each study and that the reported effectiveness of the ANN has been validated by the peer-review process. It is not within the scope of the present review to make an objective assessment for each study, and it is highly unlikely that sufficient data to support such analysis would have been provided with the individual study being assessed (few published studies include a full dataset). For example, if a peer-reviewed study claims to have provided a correct predictive measure for a certain percentage of the cases, then that claim is accepted as an accurate reflection of the ANN's success.

With regard to the second factor (validation of training), the Weight of Evidence analysis, as will be described in detail in Chapter 4 (Methodology), establishes that some measure of ANN validation is required. Various methods, such as a hold-back or a k-fold approach (Adya & Collopy, 1998; Lisboa, 2002) have been considered acceptable means of validating the training process, thus confirming that the training set created a network that is generalizable to a larger population. Each method has advantages and drawbacks, but validating samples using an historically recognized process that tests the training against untrained data offers greater confidence in the ANN application's value.

A similar approach is taken with the third factor (sample size), where a Weight of Evidence analysis can measure whether the source dataset for each study has a statistically reasonable and generally accepted sample size, as measured against established standards. Studies that relied on sample sizes below that threshold might not be acceptable for inclusion

within this study. This is in line with generally recommended statistical practice and supported by the peer-review process (see Chapter 4 for more detail on sample size). Further, it would seem appropriate that only studies that counted individual patient cases should be included; papers that counted multiple samples from single patient cases could have a confounding influence. Again, more detail is provided in Chapter 4.

Applying these quality and consistency guidelines to the selection of papers for review makes it more likely that the systematic analysis will yield reliable, relevant results about the effectiveness of ANNs.

Measures of ANN Performance.

What performance measures are appropriate to quantify a success or failure designation for the outcome of each study?

Within the context of this paper, the assessment of ANN performance consists of categorization of outcome within a dependent variable (the output) based on selected independent variables in each study's particular case (the inputs). This is a challenge since ANNs usually do not provide a clear demarcation of findings such as occurs with more traditional statistical measures (e.g., a true confidence interval is rarely provided). Zhang, Patuwo, and Hu (1998) noted that the most important measure of ANN performance is the prediction accuracy it achieves beyond the training data (p. 51), but what remains unclear is how to measure prediction accuracy in a consistent and standardized way. ANN-based research studies have been cited as using a shotgun or trial-and-error methodology, replaying the ANN building process until one achieves some satisfactory result (Zhang et al., 1998, p. 55), suggesting that as a primary reason for inconsistencies in the literature. It has also been suggested that the ANN is a learning application that can achieve high levels of performance

accuracy within training sets (as outlined in Appendix A), but the methodologies used to achieve that performance (ANN algorithm selection, number of hidden neurons, learning rate, etc.) can vary widely.

It is also true that the network can change over the course of additional training exercises. The ANN trains to the data supplied, and, with the application of further training, results may not remain static. Gant et al. (2001, p. 331) indicated that the difficulties of evaluating performance can also be expressed in ethical and legal issues that flow therefrom (suggesting that this accounts for the lack of ANN standards from national and international organizations). Yet, as mentioned in Chapter 1, the continued use of ANN in research suggests that ANNs have been accepted as effective tools, and several authors have noted the effective performance of ANNs as compared to human decision making (Baxt & Skora, 1996; Patel & Goyal, 2007).

In order to carry out this study's systematic review, it was necessary to establish a clear standard for assessing ANN performance across all of the studies selected for analysis. While there are several established approaches to assessing ANN performance, the overwhelming majority of studies examined in this systematic review (see Chapter 5) used either Prediction Percent (the percentage of correctly determined outcomes of a trained network when given non-training data) or a Receiver Operating Characteristic (ROC) curve (a graphic mapping of sensitivity versus specificity performance of a trained ANN, with maximization of both used as the performance target).³ Caruana and Niculescu-Mizil (2006, para. 2) suggested that the ROC approach seems to be preferred within the medical field. Lisboa and Taktak (2006, p. 410),

³ This is also measured by the area under the ROC curve – sometimes called AUROC or AUC (Area Under Curve); maximizing that value represents the peak ANN performance. The acronyms “ROC” and “AUROC” are used interchangeably to represent this approach.

however, noted that whether reporting the ROC or Prediction Percent, "...very few [studies] applied rigorous tests to compare their method with benchmark systems" and they suggest that performance be quantified for measurement by means of a confidence interval for a full range of ROC values (Lisboa & Taktak, 2006, p. 412).

The question remains of just how such a confidence interval could be established. Prediction Percent is the performance value assigned to a trained network based on its training set performance. It is calculated as follows: if at the end of the supervised training exercise the ANN produced the correct output for a given set of inputs 95% of the time, then the Prediction Percent is 0.95 (the expected value of its performance for non-training data). This measure focuses only upon correct predictions and does not account for Type I/II errors. One danger with this approach, of course, is that it is highly dependent upon an unbiased training set; that is, in cases where the network did not overfit to a non-representative set of training data. However, from a pure performance viewpoint, a network performing at a predictive value of 0.95 could be considered an "effective" predictor: in only 5 instances out of 100 would the network fail to accurately classify *based on what it had learned*. While Prediction Percent as a measure of success is somewhat arbitrary, it could be claimed to be no more so than a one-tailed confidence interval set at $\alpha=0.05$. Either suggests that a failure rate of 5% of the cases examined is meaningful, albeit from different perspectives. Hence, in the studies selected for this analysis, the use of Prediction Percent as a measure of ANN performance is acceptable, but for analysis of those papers that used this method it may be difficult to compare them to others that use different approaches, such as ROC curves.

In the case of ROC curves, it consists not of a single measurement but a range of values depending upon the threshold chosen for determination of success or failure. The curve itself is

a graph of the sensitivity versus the specificity of the performing ANN metric.⁴ As described by Alsing, Bauer, and Oxley (2002, p. 133), “This relation is usually used to relate the detection or ‘hit’ rate (probability of detection, i.e., probability of a true positive) to the false alarm rate (probability of false alarm, i.e., probability of false positive) as an internal decision threshold is varied”

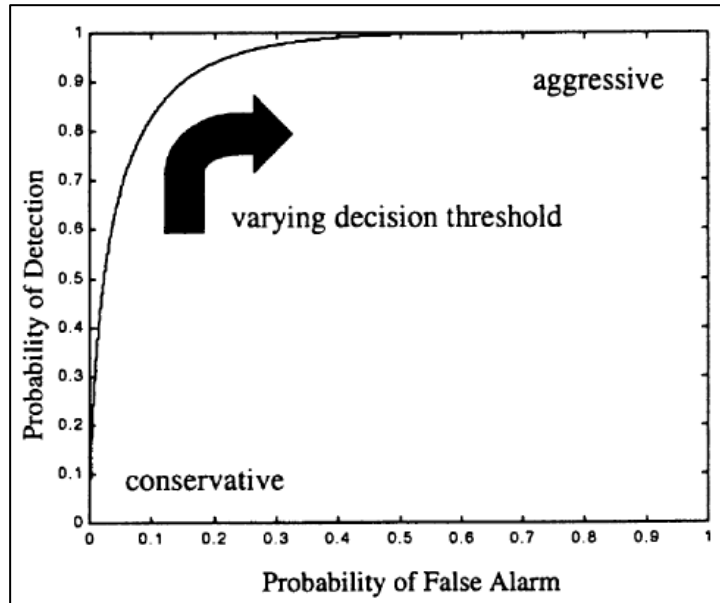


Figure 1. Typical ROC curve describing the relationship between the probabilities of detection, a true positive, versus a false alarm, a false positive (Alsing et al., 2002, p. 133, Fig. 1). Copyright 2002 by the International Journal of Smart Engineering System Design.

as schematically represented in Figure 1. When using this measurement tool DeLong, DeLong, and Clarke-Pearson (1988, p. 837) recommended using the area under the curve as an index of accuracy. They explained that “The area under the population ROC curve represents the probability that, when the variable is observed for a randomly selected individual from the abnormal population, the resulting values will be in the correct order (e.g., abnormal value higher than the normal value)” (DeLong et al., 1988, p. 837). What this does is to provide a measure in

⁴ Sensitivity is the ability to *correctly* classify, based on a specific set of criteria, an individual case as belonging to a specific category (affirming the positive), while Specificity *correctly* classifies an individual case as *not* belonging to a specific category (affirming the negative) (Parikh, Mathai, Parikh, Sekhar, & Thomas, 2008, pp. 6-7).

the range 0.50 (random performance) to 1.00 (perfect prediction), but it does not give the user any better assessment of specific performance since the curve is dependent upon the threshold chosen (itself a somewhat arbitrary measure, similar to Prediction Percent).

However, Bamber (1975) and others determined that the area under the ROC curve very nearly approximates the Mann-Whitney *U*-distribution (Bamber, 1975; Mason & Graham, 2002), thus allowing the researcher to treat this value similarly to a *t*-test statistic, and allowing the establishment of a true confidence interval to determine performance. Unfortunately, the lack of availability of the actual dataset from each study precludes us from taking on that level of analysis; there are at best only summary statistics to work with. Thus, for papers selected for this study that use a ROC, we apply the same criteria used for papers that use Prediction Percent: we accept at face value each study's findings given that every article was subjected to validation through a peer-review examination.

While the studies selected for analysis employed various methods to determine the effectiveness of ANNs, the peer review process validates those methods, and the validity assessment done for each study (supported through a Weight of Evidence analysis) would be expected to eliminate low-quality studies.

Prior Work in ANN Application Review in Healthcare

This brings us to an examination of some of the efforts done by earlier researchers in the application of ANN methodologies to the study of health and disease-related issues. A review of the literature over the last decade has identified the following six systematic reviews that focused on such ANN applications:

Studies.

- Abbod, M., Catto, J., Linkens, D., & Hamdy, F. (2007). Application of artificial intelligence to the management of urological cancer. *The Journal of Urology*, 178(4), 1150-1156. doi:10.1016/j.juro.2007.05.122
- Bartosch-Härlid, A., Andersson, B., Aho, U., Nilsson, J., & Andersson, R. (2008). Artificial neural networks in pancreatic disease. *The British Journal of Surgery*, 95(7), 817-826. doi:10.1002/bjs.6239
- Lisboa, P. J., & Taktak, A. G. (2006). The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks*, 19(4), 408-415. doi:10.1016/j.neunet.2005.10.007
- Schroder, F., & Kattan, M. (2008). The comparability of models for predicting the risk of a positive prostate biopsy with prostate-specific antigen alone: A systematic review. *European Urology*, 54(2), 274-290. doi:10.1016/j.eururo.2008.05.022
- Rajpara, S. M., Botello, A. P., Townend, J. J., & Ormerod, A. D. (2009). Systematic review of dermoscopy and digital dermoscopy/artificial intelligence for the diagnosis of melanoma. *British Journal of Dermatology*, 161(3), 591-604. doi:10.1111/j.1365-2133.2009.09093.x

Dissertation.

- Ghavami, P. K. (2012). An investigation of applications of artificial neural networks in medical prognostics. (Order No. 3542349, University of Washington). ProQuest Dissertations and Theses. Retrieved from <http://ezproxy.umuc.edu/login?url=http://search.proquest.com/docview/1197730276?accountid=14580>. (1197730276).

Each of these works has synthesized the outcomes of selected studies using similar systematic review methods employed here, but with one major exception - each limited its scope to one particular clinical issue or diagnosis, bounding their entire examination only to that realm. As mentioned earlier, this study's focus is to provide a more generalized performance assessment. The dissertation reference (Ghavami, 2012) came much closer to the intent of this paper, but the author chose to exemplify his findings through a focus on deep vein thrombosis (DVT) and pulmonary embolism (PE) patients and an ANN trained (designed) to predict disease prognosis as it relates only to those two conditions. In this study the aim is to examine ANN training across a broad spectrum of applications, and not to exemplify it via a particular model diagnostic or prognostic evaluator, hence the research questions are non-specific as to a particular outcome measure, only to predictive performance.

However, based on the collected analyses of those works above, they do provide some indication of the challenges posed by this type of systematic review study. Each of the works above were examined as to how it adapted to those challenges and limitations, and by using their lessons, where applicable, the current study was modified and enhanced accordingly (refer to Chapter 4 – Methodology for a detailed discussion of the review and analysis of this review's studies).

While the studies above were presented to demonstrate the use of systematic review in assessing ANN utilization in healthcare, there are also those for which there was specific exclusion of such methodology. Thus is provided three cases of systematic review study where computer-based algorithms and neural network applications were overtly discarded or ignored. The first, Hess et al. (2008), suggested that an in-depth review of artificial neural networks was beyond the scope of their work (p. 375) and thus ignored any studies in which they were

included. Jacob, Lewsey, Sharpin, Gimson, Rela, & van der Meulen (2005) eliminated ANN-based studies because, they suggested, it was impossible to mimic or replicate the procedure in detail (p. 815) – how the ANN actually arrived at its final configuration. These both illustrate a common problem faced by those using ANNs, and is often referred to as the “black-box” issue, as has been noted earlier. Finally, there is Nehme, Boyle, and Brown (2013) who reviewed pre-hospital care of acute coronary syndromes. Their claim was that due to, “...the limited access to technology in the prehospital care phase, it was deemed appropriate that all models requiring specialist computation were excluded from the review” (p. 952). Hence, this provides another limitation of ANN modeling – the lack of availability of ANN-based tools at the point of care, even given the data upon which to build (and train) such a model. As has been noted in Chapter 1 – Introduction, there are now several generally available statistical applications (SPSS, SAS, and MatLab, among them) that provide detachable neural network executables that can be used to port the trained ANN for use in other applications. While the complexity and ease needed to port that technology varies by the system employed to create them, the technology has seemingly reached the point that warrants, in the eyes of the ANN software creators at least, such applicability.

ANN Cost versus Value in Healthcare

One important aspect of ANN implementations in practice is that of cost versus value – that is, does the ANN tool provide sufficient value to the practice to warrant the expenditure of resources (time and money) to support its implementation and use? This general question has two aspects:

- 1) Are the costs of implementing an ANN application prohibitive, and,
- 2) Is the value that an ANN can bring to a practice sufficient to justify those costs?

Given that this paper addresses ANN use in the healthcare field, those questions will be evaluated in that context.

To address the first issue, one can look at the actual total costs of an ANN application: software, time and effort to develop a network, and the mechanism to deploy the network in practice. The development effort is certainly costly, but, as is suggested in the studies selected for this review, researchers have already assumed that burden. Whether the tool is designed to make a particular diagnosis (e.g., metastatic breast cancer) or to determine the probability of a particular prognosis (e.g., an estimated duration of inpatient rehabilitation after a spinal cord injury), each study proffers its ANN as a task-specific calculator to apply in practice. Hence, for many applications the developmental effort has already been borne by the researchers who designed and evaluated the tools. Because a market for these tools has not yet developed, the direct cost to clinical consumers is still unknown.

There are certainly costs associated with making that “calculator” available to practitioners (the second issue above). The actual costs will depend on the software product chosen by the developer (or researcher, in this case). In some software implementations a run-time version can be distributed to the user at minimal or no cost as a means of applying an established and trained ANN to a given set of case data provided by the practitioner (e.g., Ward Systems Group, Inc. – Belliveau et al., 2016). Other ANN applications would require operational licenses, for which costs can vary widely. For an institutional implementation, such as in a teaching hospital, applications may already be available for licensing (MatLab, SPSS, SAS, etc.), but a small professional practice would probably need to purchase the application. Where would one expect to find the resources required to implement the ANN tool?

In healthcare, and particularly in the not-for-profit healthcare sector, grants are often sought to support improvements in patient care, including those reliant on information technologies. However, on occasion, an institution will choose to fund a new technology directly, recognizing it as an investment in infrastructure essential to patient care. As discussed in Chapter 1, healthcare institutions typically rely on two approaches — loss/risk aversion and error reduction/mitigation — to evaluate the potential benefits of acquiring a tool like an ANN. If implementation of ANN technology can reasonably be expected to reduce risk, loss, and error, and to improve patient outcomes, without incurring costs that would exceed the benefit, then its acquisition is likely to be justified.

Next, presuming that the practitioner wishes to implement an ANN but does not have resources immediately available, how can one justify the outlay to the practice, either through capital expenditure or grant funding? A wealth of literature addresses these questions through two arguments.

The first argument speaks to the significant inefficiencies and inaccuracies inherent in healthcare practice (particularly in diagnostic assessment). The second argument is that the application of technological solutions (such as ANN tools) in healthcare practice can be expected to achieve significant reductions in time and labor costs, thereby reducing practice overhead and freeing up financial resources for other uses. In support of the first contention, a study by Graber and Carlson (2011, p. 12) claims that diagnostic error may be an enormous unmeasured cause of preventable morbidity and cost. The diagnostic error rate in clinical medicine is approximately 15% (Berner & Graber, 2008, p. S3), and mitigation of diagnostic error was suggested as the next frontier in patient safety (Newman-Toker & Pronovost, 2009). Indeed, in that previously mentioned study on diagnostic error in internal medicine (Graber, Franklin, & Gordon, 2005),

100 cases of error were reviewed to measure the severity of harm, and the analysis (p. 1494) revealed a clinical impact on the VHA scale⁵ of 3.80 ± 0.28 (Mean \pm SEM). Few studies have analyzed the cost impact associated with those errors. One study of note, however, suggests that because errors create a demand for patient services that would not otherwise be required, these errors actually increase patient hospital stays, services provided, and, perversely, associated revenues (Goode et al., 2002, p. 949). This scenario could be viewed by healthcare payers (e.g., insurers) as an opportunity for savings, and by healthcare providers (e.g., hospitals and clinical practices) as an avoidable trap.

Regarding the second factor, a report from the Institute of Medicine (2012), or IoM, which evaluated on a national scale the impact of technology on healthcare practice, acknowledged that, “Advanced statistical methods, including Bayesian analysis, allow for adaptive research designs that can learn as a study advances” (pp. 6-8). The IoM report recognized that Bayesian approaches to data (of which the ANN is an example) provide the kind of value-add healthcare needs in order to become much more effective and to reduce error rates, two improvements that will result in cost savings. However, just adding technology by itself is probably insufficient to improve practice. As noted by Lisboa (2002, p. 11) in discussing the advance of ANN tools in healthcare:

Advanced computing would enable physicians to concentrate where they are most needed, at the patient's bedside, while specialist knowledge would be left to recall systems that can handle the 'encyclopaedic' aspects of medicine (Schwartz, 1970).

However, it became apparent that the enormous complexity created by interactions

⁵ The VHA Impact Scale of Severity of Harm scores 1-4 (though actual scoring could exceed that range), with 1 being “Minor” and 4 being “Catastrophic” (Graber et al., 2005).

between clinical conditions made a comprehensive scenario analysis intractable. This started a dilemma that is still current, namely the need to specialise the design of decision support systems to closely circumscribed medical problems, when clinicians have no reason to take-up computational tools unless they are useful for almost every patient in a generic category of clinical conditions (Shortliffe, 1993).

Hence, we focus here on diagnostic and prognostic assessments of individual maladies rather than a generalized assessment of overall patient's overall well-being.

Concerning actual healthcare costs, Price Waterhouse Coopers (2010) has identified as much as \$312 billion in annual waste in clinical practice in the U.S.– \$25 billion in preventable hospital readmissions, \$17 billion in medical errors, and \$10 billion in treatment variations, all of which are implicated by errors in diagnostic or prognostic assessment. Another source suggested that, within the federal Medicare population alone, there were 238,337 patient safety-related deaths from 2004 through 2006 resulting in costs of \$8.8 billion (Jao & Hier, 2010, p. 121). Thus, there are clearly opportunities to reduce costs, not only in the larger scale, but for smaller, individual practices, as well. It is therefore warranted that examining ANN tools as a means to improve clinical assessments can be a significant contributor to that discussion.

The ANN as a Disruptive Innovation

Another way to view the use of ANNs within clinical practice is as a *disruptive innovation*, as presented in the seminal work by Christensen (1997). The ANN is *disruptive* because it introduces potentially novel information to the clinician during the decision-making process, and it is *innovative* because of the manner in which it is employed – that is, by interceding in the normal decision-making processes just at, or before, the point where the clinician achieves a final diagnostic or prognostic determination. Christensen's theory of

disruptive innovation suggests that disruptive technologies (as compared to a non-disruptive innovation) have four characteristics: (a) they are simpler, cheaper, and lower-performing; (b) they generally promise lower, not higher, profits, at least at the outset; (c) most existing customers don't use or want them; and, (d) they are first commercialized in emerging or small markets (Christensen, 1997, p. 267). However, while Christensen's context was product/service innovations relating to market and technological change (1997, p. xiii), a parallel relationship to healthcare practice may be drawn. These four elements relate to the clinical decision-making process in specific ways:

- **Simpler, cheaper, and lower-performing.** In this context, the ANN, when compared with a consulting clinician (another set of human eyes to discern the diagnostic assessment), can be reasonably assumed to be cheaper, with one software purchase covering the use across a great many patients. It is simpler and lower-performing due to its focus; the ANN has specific diagnostic/prognostic limitations, unlike the *general-purpose* human clinician who has a broad scope of diagnostic assessment capabilities.
- **The promise of lower, not higher, profits, at least initially.** The ANN provides a diagnostic tool that would, when applied, require an acquisition cost but not be directly engaged in any additional revenue generation, at least at an early juncture. Even with that apparent negative effect on practice profitability, the long-term return can be significant based on the potential for improved clinical outcomes and reduced patient costs overall, as argued above.
- **Most existing customers are not aware of ANNs and their capabilities, thus they are not in demand.** ANNs are not easily explained, defined, or understood by clinicians,

and computer-based applications for decision support, even when present, are not well-used (Berner & Graber, 2008, p. S13). ANNs are perceived as a still-developing technology that is neither understood nor accepted enough for general use (Croskerry, 2002, p. 1201). Hence, there seems to be little demand for their use in practice, as noted in Chapter 1.

- These innovations are first used in emerging or small markets. Since ANNs are not well-accepted by the general clinical community, their implementation has been almost exclusively applied in research, not clinical practice, and their general use has been considered in some sectors to be arcane (Lisboa, 2002, p. 13).

Thus, when analyzing ANN applications in relation to diagnostic or prognostic assessment, the characteristic of disruptive innovation seem to fit well. Moreover, in the disruptive innovation model, there are certain expectations about market changes (or practice changes, in the healthcare setting) that flow from Christensen's proposal, the primary one being that the disruptive technology may, at some point, replace and even dominate the market (i.e., clinical practice). Much like the stethoscope, the ANN may be found to be an equally useful innovation. The stethoscope does not require its user to have an advanced understanding of the physics of sound, only confidence that the device is effective, so it is that the ANN user need not possess advanced knowledge of neural networks and how they work, only confidence in the evidence-based foundations upon which the ANN is built. To that end, the present study becomes such a foundational effort.

Change Management and New Technologies

Finally, whenever encountering new technologies, issues relating to the management of organizational change come to the fore, and it is expected that ANN technologies would not be

an exception. Therefore, a review of the literature regarding organizational change is needed, specifically on how that may affect ANN use and adoption in practice. In the classical view of change management, it was suggested that, in order to effect a process change, the process must be unfrozen, moved to its new state, and then frozen in that state, and that resistance to such change is an expected systemic response (Lewin, 1947). From another viewpoint, change could be examined through the lens of four theories: Life-cycle, teleological, dialectic, and evolutionary (Van de Ven & Poole, 1995), for which new technologies like ANNs would likely fit the teleological model.⁶ In addition, Lewin and others contended that group dynamics played a significant role in change resistance, and that seems to have held up over time (Burnes, 2004, p. 981). However, in a more recent review of organizational change it was suggested that resistance to loss rather than change was often the driver, and that recommended strategies for addressing change resistance must be done holistically, through a multifaceted approach that includes education, communication, negotiation, and so forth (Dent & Powley, 2003). Thus, when employing a new technology, as with ANNs, both individual and group psychology dynamics need to be factored in to the implementation analysis.

When examining clinical practice change, the organizational implications would be no different. Liebowitz (2001, pp. 5-6) suggests building a supportive culture for knowledge sharing as a function of change management as a means to cultivate a buy-in effect. Liebowitz (2001, p. 6) warns that hype, over-expectations, and vaporware can terminally harm the effectiveness of that strategy. Representing those opportunities in ANN technology, much as

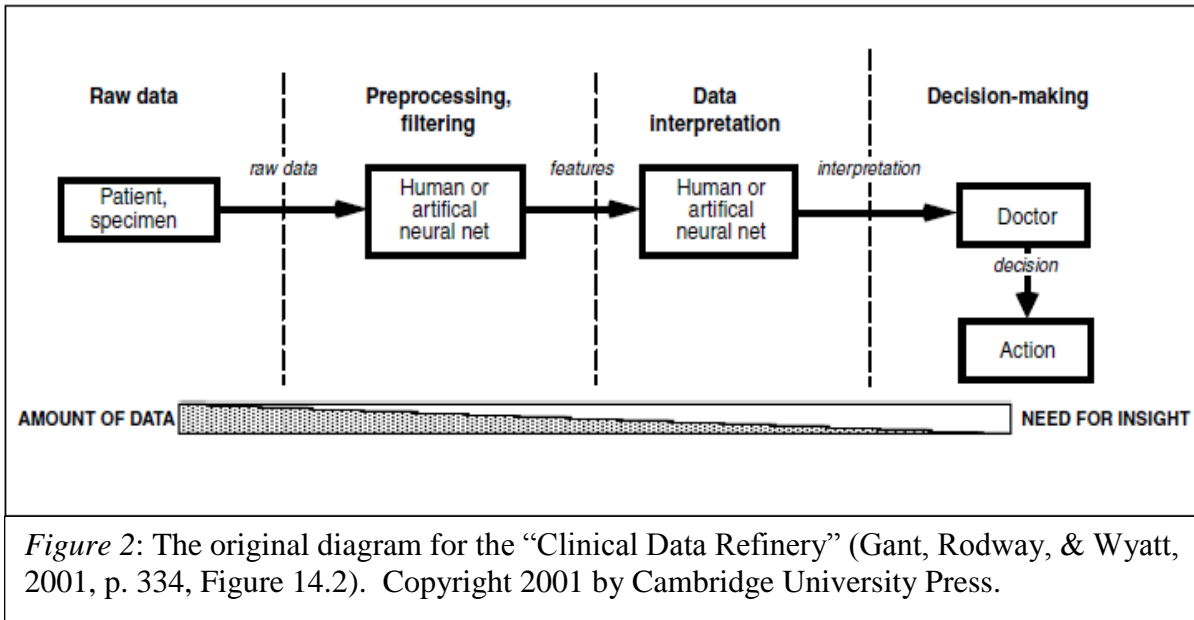
⁶ From Ven de Ven and Poole (1995, p. 535), a teleological motor is, “An individual or group exists that acts as a singular, discrete entity, which engages in reflexively monitored action to socially construct and cognitively share a common end state or goal.”

with AI technologies overall, the technologies have been lauded (Amato, López, Peña-Méndez, Vañhara, Hampl, & Havel, 2013; Coye & Kell, 2006; Makridakis, 2017; McMillan, 2017) and challenged (Amaral & Krishna, 2017; Cabitza, Rasoini, & Gensini, 2017; Marion, 2016) in the research literature as well as in news and trade publications. Thus, any theory that implements ANN technologies must include a structured warning that addresses the implementation process as it relates to organizational change.

Chapter 3 – The Conceptual Framework

Gant et al. (2001) suggested that the ANN could be a component of the clinical decision-making process that focuses and refines the initial input data into a form more readily processed by the human decision maker. As they envisioned it, the ANN could parallel human processes for preprocessing, filtering, and interpreting information, and that during the clinical decision-making process, the ANN could focus the clinician's scope of assessment. This scenario is represented in the "Data Refinery" model shown in Figure 2. In this model, the process of data assimilation (represented in the figure by "Amount of Data") decreases proportionally to the need for clinical insight, hence the value of ANNs as a tool to support or enhance human clinical decisions (Gant et al., 2001, pp. 331-334). As they described this process, "...given that there is a greater need for insight as one moves from preprocessing and filtering to interpretation and decision-making, the balance favours the human towards the right hand side [of the graphic], and leaves most opportunities open for ANN on the left" (Gant et al., 2001, p. 334). They supported this by observing that human rules become more difficult as data becomes more complex; by implication, this suggests that ANNs tend to reduce that complexity and thus contribute to better human understanding (Gant et al., 2001, p. 334). It is significant that in the Data Refinery

model, it is the clinician (the human) who formulates the final evaluation and makes the final decision.



Yet therein lies a dilemma within the Data Refinery model: that even when a human makes the ultimate decision, the process leaves open the manner in which the ANN outcome is integrated with human decision. It exposes the decision-making reliability (and frailty) of the human clinician if that clinician does not overtly recognize the value of the ANN outcome and include that outcome in the decision-making process. Indeed, human frailty in diagnostic decision-making has been well documented (Berner & Graber, 2008; Croskerry, 2002; Graber & Carlson, 2011; Graber et al., 2005; Pham, Aswani, Rosen, Lee, Huddle, Weeks, & Pronovost, 2012), as has been the challenge of considering those frailties (Croskerry, 2002; Goode et al., 2002; Smith et al., 2012). The advantage of ANN decision-making over human decision making has been suggested (Bartosch-Härlid, Andersson, Aho, Nilsson, & Andersson. 2008, p. 820), but given the lack of ANN use in practice, its superiority is not well accepted. Thus, Gant et al.’s (2001) contention that there is a gain in performance through a “data refinery” approach remains

suspect, especially given that human performance in complex decision-making is subject to cognitive errors such as confirmation bias (as explored below).

Managerial Decision Making

Whether as part of a clinical assessment or for some business or financial purpose, the effective management of the decision-making process is crucial to satisfactory outcomes. The literature is replete with such analyses and associated recommendations. Simon (1943) recognized that the task of *deciding* pervades the entire organization (p. 2). Simon also recognized that “broader rationality” could be a modifier to subconscious thought patterns that limit one’s adaptability and skill (e.g., bias), and that presentation of such rationality is a basic task of management (pp. 202-204).

Bazerman and Moore (2013) recognized the contribution of Simon’s bounded rationality theory to human thinking (pp. 5-6), as well as its extension by Thaler (2000), who Bazerman and Moore (2013, p. 11) suggested went beyond the stereotypical economic actor to include the effects of outcomes not just upon the self but upon others. In the healthcare field, such transference is an essential component of clinical practice; the benefits derived by the organization are circumscribed by the welfare of the patient. This is illustrated by the organizational tenets prescribed by the *CMS Quality Strategy 2016* report (Centers for Medicare and Medicaid Services, 2016, pp. 3-5) as a means to advance their three recommended strategic goals of better care, smarter spending, and healthier people/communities. (CMS is a unit of the U.S. Department of Health and Human Services that administers healthcare programs covering more than 100 million people – Centers for Medicare and Medicaid Services, 2017.) From Bazerman and Moore’s perspective, improving decision-making includes taking an “outsider’s view” to mitigate the overconfidence that humans have in their own decision-making prowess

(2013, p. 222-223); indeed, overconfidence has been identified as a cause of diagnostic error (Berner & Graber, 2008). The ANN can function as just such an “outsider,” providing a different perspective than the clinician’s “insider” view, one that is free from the bias potential inherent in human decision making. Some device or process is required to integrate the results of the ANN “outsider” with the skillset and experience of the clinician “insider,” and Thaler and Sunstein (2009) provide just such a theoretical mechanism.

The Nudge

The intercession of an ANN tool at the clinician’s point of decision can be conceptualized as a *nudge*. Thaler and Sunstein (2009) defined a nudge as, “...any aspect of the choice architecture that alters people’s behavior in a predictable way without forbidding any options or significantly changing their...incentives” (p. 6). They identified several situations where nudges would be appropriate: “For decisions that are difficult and rare, when the decision makers do not get prompt feedback, and when they have trouble translating aspects of the situation into terms they can easily understand” (Thaler & Sunstein, 2009, p. 74). This suggests that the nudge is neither decisive nor definitive, and that the success of the nudge relies on human recognition of its value at the time the nudge is provided. The point where the ANN intersects with the clinical assessment is just such an opportunity, where the diagnostic/prognostic decision still relies upon the clinician’s skilled judgment, as well as the recognition by the clinician of the value that the nudge could bring to the decision. This opportunity in clinical decision making was recognized by Lisboa and Taktak (2006, p. 413), who, in their systematic review of ANNs in cancer research, noted, “...computerized decision support systems serve not to instruct on a decision on a predicted outcome, but to modulate the clinician’s own decision by adding new evidence

through associative recall from historical data.” It is just this opportunity which the *nudge*, as described above, provides.

Concerning clinical assessment, certainly the process is not an easy one as demonstrated by the degree of expertise and experience required to attain the requisite skill (with the clinician typically achieving a doctorate degree along with years, if not decades, of experience in the field). While some clinical assessments may be straightforward and repeatable processes, such as those determined through targeted laboratory tests, there is ample evidence that assessment and diagnoses are complex processes prone to human error. The Institute of Medicine’s (1999) seminal report *To Err is Human* suggested that as many as 98,000 people die each year due to medical error, including misdiagnosis, implying that current practice is significantly flawed. Much more recently, in a six-year retrospective review of clinical diagnoses through autopsy findings, almost 10% of the 334 cases examined were identified as “class I” discrepancies – the autopsy revealing a diagnostic error with a potential impact on survival or treatment, including untreated infection, pulmonary embolism, and undiagnosed malignancy (Marshall & Milikowski, 2017). This suggests that some of those problems identified over 15 years ago are still unabated in clinical practice.

Interestingly, in another study, which focused on diagnostic error, it was determined that, of the 100 cases of diagnostic error reviewed, 74 involved some form of cognitive error, mostly due to faulty synthesis or faulty data gathering – but only 11 were due to faulty knowledge or the lack of requisite skills (Graber et al., 2005, p. 1495). This suggests significant cognitive process involvement in the diagnostic decision errors while less so for skills and competency – the diagnosing clinicians are capable, but they face challenges within the processes they use to come up with a proper diagnostic assessment. And, as has been discussed, human cognitive biases and

heuristic thinking can influence the decision-making process even of the most expert clinician, as recognized in a more recent IoM report on diagnostics in healthcare (Balogh, Miller, & Ball, 2015, pp. 31-68). The factors of complexity, difficulty, and human bias, make clinical assessment a candidate application for an ANN's objective "nudge."

Among Thaler and Sunstein's (2009) list of situations where a nudge from an ANN would be appropriate, they cited instances "when the decision makers do not get prompt feedback" (p. 74). This lack of "feedback" can be interpreted to include situations where critical information (e.g., results of lab and radiology tests, outcomes from prior treatments) has been delayed or is not available at the time of a patient visit, especially an initial visit to a new provider. Follow-on therapist evaluation and treatment orders can take several sessions to complete, and days or weeks can elapse before a significant outcome determination is reached and reported. While laboratory and radiology orders tend to be processed more quickly, even those processes can be protracted due to a variety of reasons, including the ancillary service provider's resource availability and characteristics of inter-institutional communication and reporting systems, as suggested by Hawkins (2007). Hawkins further observed that even in hospital Emergency Departments (EDs), the typical turnaround time for lab tests is as much as one to two hours (Hawkins, 2007, pp. 184-185), though clinical staff need lab results much sooner to support effective decision-making (pp. 181-182). Delivery of lab results to non-ED clinics or medical offices typically takes even longer, due to distance from laboratories and the extra time needed to transfer specimens and communicate results. Thus, many factors can contribute to the delay of timely feedback after a patient's initial presentation, which can lead to significant expenditures of time and resources in preliminary diagnostic assessments and patient interventions, both of which call for even greater reliance upon the clinician's assessment skill.

Thus, as Thaler and Sunstein (2009) suggested, any tool that can provide a subtle, objective, evidence-based “nudge” to the clinician at the point of assessment would tend to improve the clinician’s performance. As Thaler and Sunstein (2009) stated, “An important type of feedback is a warning that things are going wrong, or, even more helpful, are about to go wrong” (p. 92). Indeed, clinicians are familiar with this sort of feedback, as part of their practice is to “nudge” their patients to take steps to improve their own healthcare outcomes (Blumenthal-Barby & Burroughs, 2012). In similar fashion, at critical decision points clinicians could benefit from the delivery of a parallel, evidence-based machine-learning derived assessment, whether confirmatory or contradictory. Therefore a newer model beyond the one posited by Gant et al. (2001) is required to integrate these theories into one that can be implemented into clinical practice.

The Concept Model Proposition

As noted, Gant et al.’s (2001) original model did not include some established theories about the human decision-making process. In his dissertation, Herbert Simon (1943) suggested that the human decision-making process incorporates subconscious patterns that can limit the decision-maker’s adaptability and skill (p. 202). Humans, Simon said, are not automata that grind out decisions in clear, predictable ways. Rather, the decision process is complex and subject to external influences that, as Simon (1943, p. 202) suggested, can be rationally modified when they are recognized and understood. This concept is popularized in the current literature as the *Homo economicus versus Homo sapiens* debate (Thaler, 2000). Several theoretical models explain this concept; three of them, chosen for inclusion in this study, can be viewed as illustrative of many of the cognitive behaviors involved in human decision making:

- 1) Satisficing – As defined by Simon (1997), the administrator (i.e., the decision-maker) looks for a course of action that is to some degree satisfactory or “good enough” (p. 119). “Good enough,” however, is a discretionary judgment based upon data availability and search sufficiency. From a decision-making point of view, the benefit – and risk – of satisficing is that when the goal is to expedite a decision, the process can truncate the flow of information and discourage the evaluative thought process (Bate, Hutchinson, Underhill, and Maskrey, 2012, p. 615). As posited by Campitelli and Gobet (2010), using the satisficing process implies relying on experience to define the expectation of how good the solution needs to be, and to stop the process when “good enough” has been achieved (p. 361).
- 2) Sensemaking – In his seminal work, Weick (1993) posited that in organizations, “the basic idea of sensemaking is that reality is an ongoing accomplishment that emerges from efforts to create order and make retrospective sense of what occurs” (p. 635). In the healthcare environment, the sensemaking process might be exemplified by how an Emergency Department physician responds to a large-scale trauma event. The physician must triage and manage several, possibly many, critical cases simultaneously, while remaining clear-headed and objective enough to conduct assessments effectively without succumbing to distractions or being overwhelmed by the stress of being exposed to trauma in a medically urgent environment, which itself can lead to overconfidence (Berner & Graber, 2008). Despite physicians’ generally well-developed metacognitive skills, Berner and Graber (2008, p. 519) noted that it is certainty of one’s skills, not uncertainty, which seems to contribute most strongly to diagnostic error. That is, the more confident the clinician is in his or her assessment,

the greater the probability of diagnostic error. And, as Weick (1993) noted, “Extreme confidence and extreme caution both can destroy what organizations most need in changing times, namely, curiosity, openness, and complex sensing” (p. 641). Further, under crisis conditions, exposure to trauma can have adverse effect on human cognitive processes (Pearson & Clair, 1998; Weick, 2005).

- 3) Heuristic thinking – Heuristic thinking (e.g., applying rules of thumb, making educated guesses), while intrinsic to the diagnostic process, has been found to be problematic in clinical practice. Heuristic thinking is a natural consequence of critical thinking, which is a key component of clinical practice (West, Toplak, & Stanovich, 2008). As Simon (1979) recognized, heuristics seem to be a guiding principle in the psychology of problem-solving (p. 507). Tversky and Kahneman (1974) also recognized this in a precursor to their landmark work on prospect theory. However, when clinicians apply rules of thumb (heuristic thinking) to patient diagnostic or prognostic decisions, those rules were sometimes found insufficient to address the complexity of the individual cases, and even more so when the clinician was confident of his or her initial decision. Researchers have documented these flaws while noting that heuristic thinking is an established and accepted practice in clinical assessment (Croskerry, 2002; Croskerry, 2014; Graber et al., 2005; Kahneman, 2011; Simon, 1992; Stanovich & West, 2000; Thammasitboon & Cutrer, 2013).

Given these cognitive models as mediators to the clinical decision-making process, this paper proposes a variant of Gant et al.’s (2001) Data Refinery model as its conceptual framework. In this model, the output of the ANN network functions as a “nudge” (as posited by Thaler & Sunstein, 2009) to the clinician - the Data Refinery using an ANN with a Nudge

(hereafter referred to as DRAWN); the nudge can either confirm or contradict the clinician's assessment. This variant model, shown graphically as Figure 3, fulfills the proposition by Lisboa and Taktak (2006, p. 413) that the ANN modulate the clinician's decision, and it is described as follows. At the outset, a validated ANN application specific to the diagnostic/prognostic question is engaged (this process, represented by the ANN Network operation shown in Figure 3, will be discussed in more detail below). Input to the ANN consists of the particular case of concern in order to classify that case; the output (ANN Outcome) is presented to the clinician as a recommendation. The network does not supersede, but rather augments, the human decision-making process, and the patient case remains available to the clinician for review during the actual clinical assessment and subsequent clinical interventions.

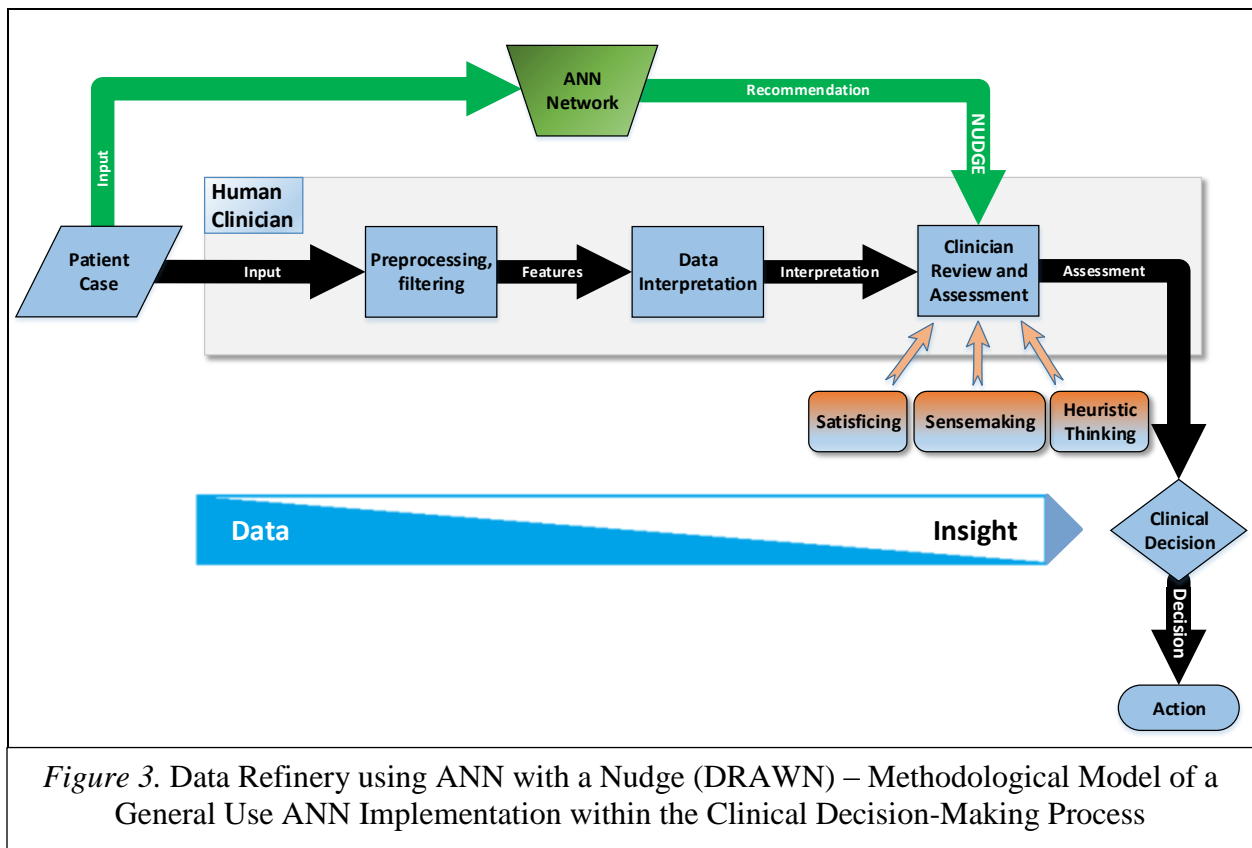
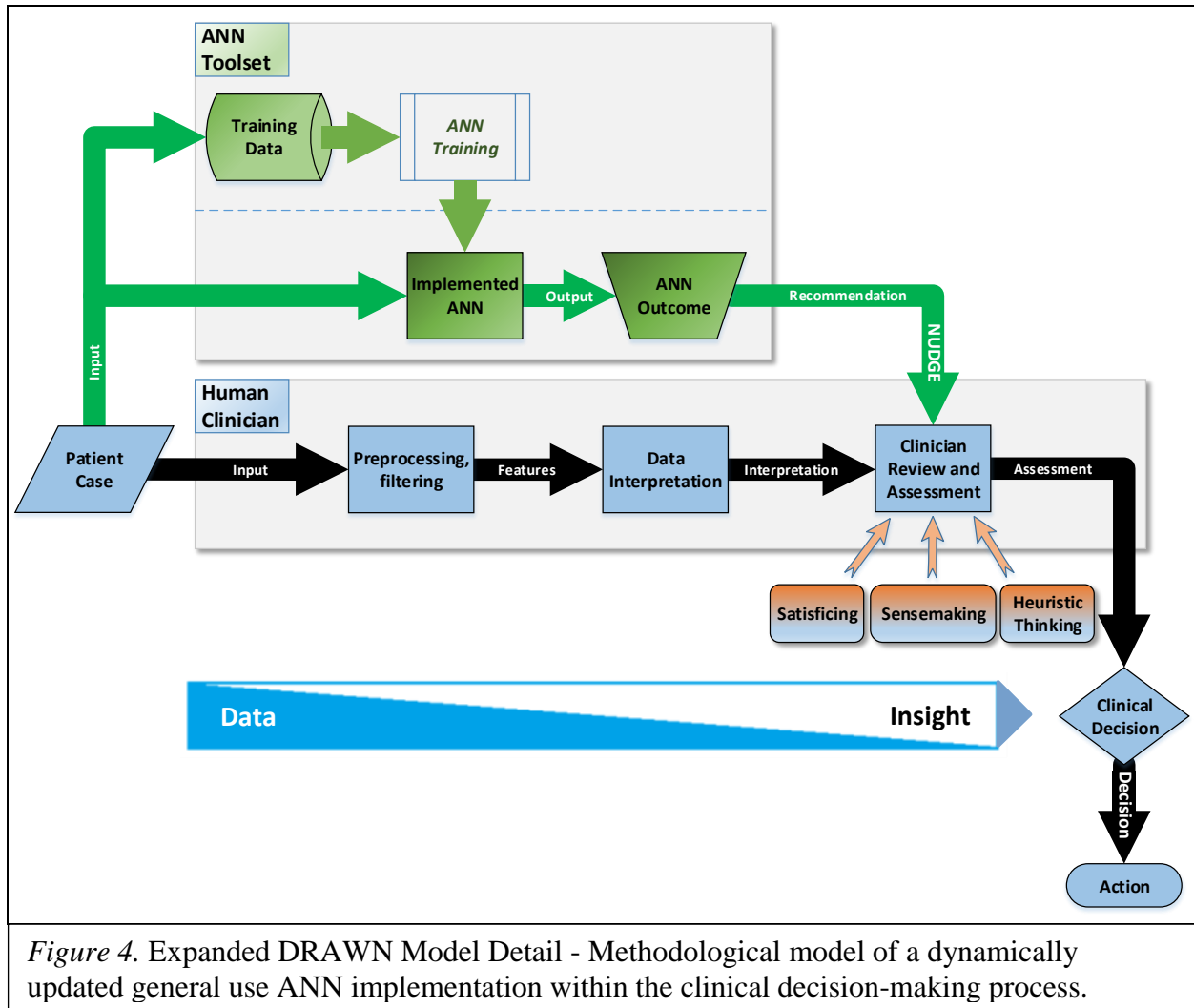


Figure 3. Data Refinery using ANN with a Nudge (DRAWN) – Methodological Model of a General Use ANN Implementation within the Clinical Decision-Making Process

However, the clinical assessment is itself a process of significant complexity, as noted above and as illustrated in Figure 3. The cognitive process theories discussed above (and possibly others), can influence the human decision-making process. These potential influences are modeled here as *mediator* variables that influence the outcome, as opposed to *moderators* which are more directive towards it; the variables are based upon the theoretical processes that suggest how humans arrive at decisions.

One further aspect of this approach warrants discussion, namely, the post-analysis inclusion of the newly evaluated patient case back into the ANN training dataset. The introduction of this new data refines the ANN tool and maximizes the value of findings for future cases. Thus, the ANN is iterative and can be continuously developed and expanded, rather than existing solely as a static model as would be expected for a typical mathematical archetype (e.g., multivariate regression). This final step is shown in the complete version of the model as represented as Figure 4, where the *ANN Network* model shown in Figure 3 has been expanded to include the training toolset. In this model, the ANN becomes an ongoing learning toolset with the expectation that accuracy will continue to improve as more cases are added to the training dataset. In this way, the ANN “learns” from its experience, similarly to how humans improve decision-making skills through continued experience.



This model of evidence-based, ANN-augmented, clinical decision-making becomes the foundation for this proposal for a new approach to clinical assessment for diagnostic and prognostic purposes. However, it is first necessary to demonstrate that ANN tools are generally sound, effective devices for supporting clinical assessments, hence the object of this systematic review.

The Venn diagram in Figure 5 illustrates the use of ANN technology in clinical studies across a spectrum of applications, suggesting its potential contribution to clinical decision making. This is not to suggest that the ANN algorithms across these implementations are

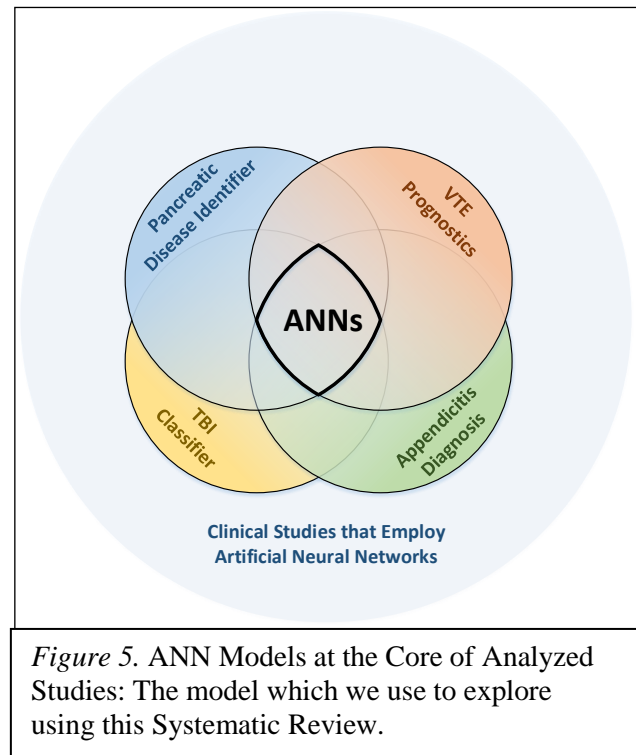
identical, but that the functional usage is the same: that is, that a learning system is applied to training data that configures itself to the relationships found within the data and makes its findings available for use in non-training cases.

Understanding the Potential for ANN Use in Practice

Given the presented definition and description of the ANN, its function, and its potential application to reliable healthcare decision making, this brings us back to the question noted earlier: Why bother? Why should healthcare institutions invest in ANN tools if they can do nothing more confirm the knowledge, expertise, and experience of highly skilled clinicians?

One answer lies in how in the context of individual cases the ANN tool can raise issues or questions not previously recognized or known. This is due to how the ANN learns – not as humans do, from limited personal experience or a finite set of prescriptive texts and lessons – but from the non-presumptive dataset of evidenced patient parameters and validated outcomes.

In addition, ANNs do not display any of the cognitive flaws inherent in human decision making such as confirmation bias (that is, the expectation that subsequent evidence is of particular value only when it is in support of initial findings – Rousseau, 2012, p. 6). Indeed,



Flyvbjerg (2004), in his review of case study research, observed that Francis Bacon recognized this attribute of human decision making as early as 1852:

The human understanding from its peculiar nature, easily supposes a greater degree of order and equality in things than it really finds. When any proposition has been laid down, the human understanding forces everything else to add fresh support and confirmation. It is the peculiar and perpetual error of the human understanding to be more moved and excited by affirmatives than negatives. (p. 428).

By contrast, the evidence-driven ANN has no confirmation reference, only the objective mapping of inputs to output variables as learned from the training set. When presented with contradictory information, the ANN simply uses that data to adjust its network as necessary to reflect that discrepancy, unlike humans, who tend to ignore contradictions that do not align with their expectations.

A theory of human thinking that contributes to this understanding was recently introduced by Nobel laureate Daniel Kahneman: WYSIATI, or What You See Is All There Is (Kahneman, 2011, p. 85). Kahneman, who defined WYSIATI as “Jumping to conclusions on the basis of limited evidence,” considered this concept key to understanding intuitive thinking in humans (Kahneman, 2011, p. 86). The concept is supported by other well-accepted research (Pettigrew & Roberts, 2006, p. 130). Within the context of the ANN tool, generating the network with insufficient or incomplete data could limit effective decision-making in the same way that jumping to a conclusion can do for a human decision maker. However, our review has accounted for this through the Weight of Evidence analysis (see Chapter 4) where we reject studies that rely on insufficient data.

A related aspect of Kahneman's work was the research that he and Amos Tversky did on what they called the *availability heuristic* – that is, the process of judging a decision by the ease with which seemingly similar instances come to mind (Kahneman, 2011, p. 129). This can be related to how one might design a categorization study, defining subjectively what groupings to analyze and deciding how to stratify the data. Are the chosen categories the best match for what actually appears in the data, or were investigators' decisions biased by (for example) clinical practice standards, or even by their reading of a recent study that used a similar approach to gradation of data? Though the choices we make may at first seem clear and reasonable, with greater scrutiny they might not appear as objective or as effective as some alternatives; yet those choices will have a potentially significant effect on the outcome and findings of the study.

The ANN, on the other hand, has no preset conditions, no availability heuristic to use as a resource (unless we use a pre-loaded/pre-configured network to start, rather than trained one – a very unwise choice, for good reason as noted here). In addition, there is no preliminary finding from which to assert a confirmation bias or availability heuristic. The ANN is influenced only by the plodding iterative analysis of the training dataset (and later additions of clinical cases, as mentioned above), which likely includes information not easily available within the clinician's knowledge base. This does not imply that ANNs are without their potential pitfalls, such as noted by the local minimum (gradient descent algorithm) issue presented in Appendix A. However, given the mitigating approaches derived through the network training process – such as use of a sufficiently-sized training data set and determination of network reasonableness by using specific test cases for validation – the ANN should provide a reliable and effective tool in establishing relationships between the network inputs and its output(s), based solely upon its training data set.

From the perspective of the healthcare manager, this effectiveness can be examined in the context of organizational change, specifically change management. Social scientists Van de Ven and Poole (1995) viewed change as involving a connected sequence of events, decisions and actions. What differs here is the degree to which the change theory they present follows certain essential stages and the extent to which the direction of change is constructed or predetermined. As noted earlier, Van de Ven and Poole posited four “ideal-type” theories, or “motors” that drive social and biological change: life-cycle, teleological, dialectical, and evolutionary (Van de Ven & Poole, 1995, p. 511-513). While they describe some ordination to those related events, decisions, and actions, the concept is not necessarily restricted to a temporal frame of immediacy, as one might suggest is the case for a clinical assessment within the context of a clinical practice. Thus, the concept model proposed here suggests the application of Thaler's decision-making *nudge* (Thaler, 2000; Thaler & Sunstein, 2009). The ANN is a decision influencer, but its timing is crucial to the decision and how it is made. If the ANN is presented too early in the process, it becomes, essentially, a fall guy for the decision-making process (in Simon's, 1979, terms, it *satisfices* the requirement for the decision, hence no further action is required). If it is presented too late in the process, then there is opportunity for human confirmation bias to intervene (Kahneman, 2011; Mendel, et al., 2011), enabling the decision-maker to resist any explanation that does not conform to his or her established thought. Conceptually similar to Malcolm Gladwell's (2002) concept of a *tipping point*, the ANN represents a temporally critical intervention that may, or may not, alter a process.

Dynamic Clinical Decision-making Logic Chain

As a means of summarizing the theoretical propositions laid out by this chapter, a logic chain table is presented in Appendix E to illustrate how the referenced social science theories

described above play into the design of the concept model presented earlier in Figure 4. The basis of that table is found in the following seven proposition statements:

- Humans have been shown to be flawed decision makers.
- Cognitive mechanisms have been shown to contribute to human analytical limitations.
- ANNs are evidence-based decision networks based on collections of relevant data to which the ANN has been trained, and which are commonly applied in clinical research.
- ANNs have been used almost exclusively in research, but only rarely applied to clinical practice.
- Employing an ANN as a “nudge” within the clinical decision-making process can prove effective in improving decision outcomes.
- The cost of ANN implementation can be offset by a reduction in clinical judgement error and its associated costs.
- The use of ANNs in research has yielded evidence for their value to practice, showing them to be effective decision-making tools (the primary research contribution of this paper).

As the table indicates, each of the seven propositions is described in column two, and the theoretical basis for those claims is presented in the column three. This table, then, presents the evidentiary basis for the proposed concept model, and provides the foundation upon which any assertions proposed in Chapter 6, along with this study’s results in Chapter 5, might communicate a measure of the value this research provides to clinical practice.

Chapter 4 – Methodology

After the discussion on ANN technology (Appendix A), its use in published clinical research (Chapter 2), and the conceptual framework (Chapter 3), what follows is a review of the methodology employed for this systematic review process, specifically an examination of the search protocols used. But before that discussion there are some research question parameters that must be explained – the selection of only those studies that use ANNs as prognostic and diagnostic classifiers, and a discussion of how ANNs in comparative methodology studies (those that used ANNs along with other techniques and compared their findings) were analyzed.

To begin, the research question analysis is limited to those works where the ANN was used as a classifier and not as a generalized predictor, time-series forecaster, or clustering analyzer (as defined by Bigus, 1996, pp. 38-40). The reason for this selection is that classifiers have been identified for their particular use in healthcare applications, specifically for prognostic and diagnostic studies (Amato et al., 2013; Ghavami, 2012; Kumar & Kumar, 2013). Further, such applications of ANNs are directly employable within a clinical practice, where diagnostic and prognostic classification is essential to the determination of patient treatment. As the intent here is to determine if ANN usage in practice is warranted, then it is appropriate to limit that examination to those applications where use in practice is relevant and appropriate.

Next we need to define how those studies which used ANN-based tools as one of several concurrent approaches (versus exclusively) were dealt with in this review, and how those components were isolated for inclusion within the study. As part of this systematic review, we retained studies that contained a performance measure for a diagnostic/prognostic assessment process using an ANN regardless of whether that was the only application evaluated by the source study. This is supported by an accepted purpose of systematic review, namely, that it is

intended to produce more than a summary of the primary studies (Briner & Denyer, 2012, p. 123; Briner, Denyer, & Rousseau, 2009, p. 26). Thus, if a study compared ANNs and one or more alternate methods of assessment (for example, a pre-defined clinical algorithm in general use) the study would still be relevant, but only as to its ANN components; references to, and comparisons with, other methods and approaches would be disregarded for this analysis.

Initial Search Protocol Design

An initial search protocol was created to identify healthcare-related studies that used ANN systems as the principal means of evaluating a diagnostic or prognostic assessment, specifically using ANN as a classification tool to select among various outcomes. Depending on the ANN model, the classification would be determined by the tool's outcome variable(s) with the inputs sourced from the dataset under analysis by the study, all processed using the software that study selected. Of particular interest was how the studies identified their outcomes in terms of performance, or effectiveness, as rated by some measure of the accuracy of the ANN predictions.

As was noted in Chapter 2, many of the studies used one of two techniques for assessing performance. The first technique is *prediction rate*, that is, the number of correct predictions as a percent of all predictions made; this reports performance as a rate of prediction success calculated from the network derived from the training dataset, and representing the general success rate of the final ANN form. If the study's training dataset were representative of the full population, then this rate should not vary greatly from that of the general population. The other technique, the Receiver Operating Characteristic (ROC), is the graphical representation of the trade-off between the network's *sensitivity* and its *specificity* (also defined in Chapter 2), and what has been called a "threshold" – that is, the decision point that determines success or failure

for the output (Faisal, Taib, & Ibrahim, 2012, pp. 665-666; Holst, Ohlsson, Peterson, & Edenbrandt, 1999, p. 414; Lee et al., 2010, p. 1480; Uğuz, 2012, p. 70). As with prediction rate, given a representative training set the ROC measure should be generalizable. Independently reported sensitivity/specificity measures can be viewed similarly, given that ROC values themselves are directly based upon sensitivity and specificity measures.

While prediction rate and ROC (and, consequently, sensitivity/specificity), as used with ANNs, are nonparametric measures of performance (there is no underlying assumption made about data distribution, and large datasets are not summarized prior to assessment – Russell & Norvig, 2014, p. 769), they are predicated on evaluating the ANN's ability to predict real-world outcomes. As noted earlier, we have assumed that the success of ANN implementations as reported in these studies has been confirmed through the peer-review process. It remains to be demonstrated whether any particular outcome prediction rate is materially better than any other (for example, how does a rate of 88.5% compare to one of 84.3%). We have only the presumption that those are both substantially better than 50% (that of pure chance); indeed, this remains a primary challenge associated with the application of ANN technology to clinical decision making, one that may be resolved over time as more research is carried out and as real-life applications can be evaluated. Nonetheless, the findings of the reviewed studies will stand as the yardstick of success within the present paper.

Information Sources and the Boolean Search Criteria

Next, we review the actual search protocol used to identify relevant studies. As with other phases of this study, the initial search activities provided insights that promoted revision of search criteria, as identified below. The entire process has been summarized in Appendix F (PRISMA 2009 Flow Diagram), in accordance with the PRISMA Statement (Liberati et al,

2009). A detailed description of each of the steps is given below and in the following two sections of this chapter.

The initial search was completed in October 2016 in UMUC Library's OneSearch database, a service provided to UMUC students by contract with EBSCO Information Services. UMUC OneSearch allows simultaneous searches of databases to which the UMUC Library subscribes, which at the time of the search completion included 53 databases and seven "additional resources" (UMUC Library, 2016, Which databases are included in OneSearch?).

The Boolean criteria for that search selection were as follows:

- Subject: ("health*" OR "medic*" OR "clinic*")
- All: "artificial neural network" AND "classif*"
- Include: (Full-Text / Peer-reviewed)
- Timeframe: >=2004 (keeping to current studies, defined here as those within the decade prior to the initial search year of 2014).
- Source databases: UMUC OneSource (EBSCO), MEDLINE, CINAHL Complete

This search yielded 64 published articles. In consideration of the desired breadth of this study, this initial set was augmented by additional searches from two other sources, iCONN (OneSearch at the University of Connecticut in Storrs, CT) and the National Library of Medicine/National Institutes of Health's PubMed database. Because of differences in the user interfaces between these resources and the UMUC OneSearch library application (the Subject selections did not map to the search engine's requisite topics), the output yielded responses in the thousands. After refining the output by sorting for "Relevance," 100 articles (of roughly 11,000) were selected from iCONN and 200 (of roughly 17,000) from NLM/NIH PubMed. Concerning

gray, or unpublished, research, an extensive search yielded only one item, but it did not meet the initial selection criteria and so was not included.

Some papers identified for this study were available only as pre-publication authors' manuscripts; access to the published versions was possible only on a fee or subscription basis. This was particularly true for some of the articles retrieved through the NLM/NIH PubMed database. The National Institutes of Health (NIH) Public Access database provided author manuscripts whereas other resources provided only abstracts. According to NLM/NIH PubMed Central (2017), an author manuscript is a version of a paper that has been peer reviewed and accepted for publication by the journal, and the only remaining steps between it and the published journal are copyediting and typesetting (Author Manuscripts in PMC, para. 2). Therefore, for the purpose of this study, these authors' manuscripts are considered to be equivalent to their published versions in quality and academic value.

Together, these search processes yielded citations for 364 candidate articles whose citations were imported into a Microsoft Access database for analysis. Each citation was assigned a unique Study ID number (1-364) to aid in cross-referencing during the latter steps of the review. After identification and removal of 75 duplicate citations (identified as articles having in common both the title and primary author), 289 articles remained in the dataset. Next, the abstracts were reviewed to identify articles where ANNs were used for a classification analysis (clinical diagnosis or prognosis) and in which ANNs were a primary focus of the research, that were related to healthcare or medicine.

Abstract Review and Analysis

These 289 abstracts were screened twice, first to identify studies that focused on diagnosis or prognosis, and the second to identify those that were germane to this systematic

review. In the first screening, each abstract was reviewed to determine if the study included either a diagnostic or prognostic outcome; that is, was the study done in order to determine the presence of a particular disease, malady, or set of symptoms (diagnostic) or was it done in order to predict future outcomes of an existing disease or malady (prognostic)? The abstract assessment was a brief examination based only on the abstract's content, while the detailed diagnostic/prognostic determination is a part of the Weight of Evidence appraisal for studies that passed this initial screening process. These two measures relate very strongly to direct clinical practice and patient care, and are not theoretical assessments, such as attempting to isolate a set of gene expressions or classifying clinical device performance efficiencies. The abstracts were also examined to identify studies specific to human care and outcomes and to eliminate those focused on animal or plant topics.

In the second screening, each abstract was reviewed to determine if the study was germane to this systematic review; that is, did the abstract suggest that the study was, indeed, primarily an evaluation of ANN performance (possibly in conjunction with other methodologies), and was the study clearly related to healthcare clinical practice? This phase of the review identified several *false positives*, or articles that matched the search criteria but turned out to be not relevant. For example, several studies were drug performance assessments while others focused on pattern recognition not related to diagnostic/prognostic assessment. These were removed from the dataset.

Thus, 162 of the 289 studies did not survive the abstract screening process, leaving 127 studies that would be subjected to a Weight of Evidence appraisal. The references for these 127 studies are listed separately in Appendix G.

Weight of Evidence Analysis

The methodology employed to assess the fitness of the remaining studies was proposed by Gough (2007) in his framework for appraisal of quality and relevance of evidence, referred to as Weight of Evidence (WoE). As Gough stated (2007, p. 214), “Being specific about what we know and how we know it requires us to become clearer about the nature of the evaluative judgments we are making about the questions that we are asking, the evidence we select, and the manner in which we appraise and use it.” Later, with Harden, he recommended WoE as a, “...framework [that] can be used as a practical strategy for critical appraisal to ensure that all...dimensions are systematically considered in a review” (Harden & Gough, 2012, p. 161). Gough (2007, p. 223) also suggested that the WoE framework should divide the evidence among four categories:

- A) A generic judgment about the *coherence and integrity of the evidence* in its own terms.
- B) A review-specific judgment about the *appropriateness of that form of evidence* for answering the review question (the “fit for purpose” assessment)
- C) A review-specific judgment about the *relevance of the focus of the evidence* for the review question.
- D) An overall assessment of the *extent that a study contributes evidence* to answering the review question.

To that end, a rubric based on Gough’s WoE framework (Table 4-1) was developed by which to quantify the worthiness of each study and its applicability to this review, specifically its research questions.

Table 4-1.

Weight of Evidence Appraisal Scoring

Weight of Evidence	Description	High	Medium	Low
A	Quality of study execution Clarity of purpose: accuracy; understandability; method-specific quality.	3	2	1
B	Appropriateness, relates to the review question Fit for methodological purpose.	3	2	1
C	Provides a relevant answer to the research question Has utility and value, findings were generated ethically and appropriately.	3	2	1
D	Overall assessment of weight of evidence findings A mathematical sum of A through C above.	8-9	5-7	3-4

Note. Weight of Evidence appraisal scoring matrix adapted from Gough, 2007, and from Harden & Gough, 2012.

A set of questions relating to each WoE category was developed to uniformly evaluate the WoE scoring of each of the papers under review. The questions were used within each of the WoE categorizations below and replicated in Appendix H (Weight of Evidence Scoring Matrix). In that matrix, each study is shown in the first two columns, indicating its Study ID number and the article's author and date of publication. These are followed by the WoE scoring assessments which are described below.

Weight of Evidence Analysis Questions

1. Weight of Evidence – A (Quality of execution)
 - a. Was the number of cases examined sufficient (at least 30 were required to ensure some degree of statistical significance)? That 30 case minimum was determined by Hogg, Tanis, and Zimmerman (2015, p. 202) to be a rule of thumb

measure based on a normally distributed sample. VanVoorhis and Morgan (2007, p. 48) suggested around 50 for regression-related studies (similar to classification ANNs), but in order not to exclude articles that employed a statistical process (ANNs) that were not well studied for case minimums, the lower measure was selected for use here. In addition, only individual patient case selections were counted since multiple samples from the same patient could introduce confounding issues.

- b. Was a defined ANN validation process employed? (This would typically consist of either a k-fold cross analysis of the data, or at least a hold-back; for example, training done on 30% of the data, validation on another 30%, and the remaining 40% used for actual testing, or studies that split the dataset between just training and testing)
- c. Was the study clear in its introduction *and* in its outcomes as being appropriate to either a diagnostic or a prognostic application?

Weight of Evidence – B (Appropriateness / fit for methodological purpose)

- a. Was the study focused solely on ANN performance? Or did ANN performance make up a major component of the analysis (for example, a comparison between ANN and Logistic Regression)? (Recall that we earlier acknowledged accepting comparison studies that included ANNs among the candidate tools being evaluated.)

Weight of Evidence – C (Provides a relevant answer to this study's RQs)

- a. Was ANN performance outcome measured and provided in a standardized form, such as Prediction Percent, Specificity/Sensitivity Scores, or as an Area Under the Receiver Operating Curve [AUROC or ROC] Score)?

- b. Were sufficient details provided for the ANN technology examined, including any hybridization or special algorithmic components applied, as required for RQ #2)?

The “Accept?” column contains the WoE inclusion determination (Yes/No) for each study. The “Comments” column records observations about papers that did not fit the model as required for this study (for example, studies that employ Complex Discrete Wavelet Functions – highly complex variants of ANNs – that would not allow for the same “footing” to be established as for more traditional ANN implementations, such as a Multilayer Perceptron Neural Network [MLPNN] form). The WoE assessment was done using QSR International’s NVivo 11 Pro for Windows (version 11.4.1.1064, 64-bit); This enabled us to directly connect the response for each question to a specific line, paragraph, or section within the study being evaluated, both to establish coding transparency and to provide a method for cross-validation should there be challenges to the methodology employed.

The question assessment responses are noted in Appendix F (Weight of Evidence Analysis), with each of the question-related column’s entries having one of three possible values:

- Yes – the question was fully satisfied by this study.
- Some – Some aspects of the question were satisfied, but the question was only partially addressed.
- No – the question was not satisfied by this study.

A subjective assessment was then made for the WoE category, and a score assigned and entered into the appropriate “Score” column, using a High (3), Medium (2), or Low (1) compliance value for each W-E category. The scores were totaled (WoE-A through WoE-C) to arrive at a WoE-D score (“Total”) which was then used for the final determination of inclusion.

Finally, full texts of eight of the 127 target articles were not available for review due to licensing restraints of the search database used; this was noted in the comments for each study.

Final WoE Determination

Each of the studies was reviewed and scored in accordance with each WoE category (A-C), and those results were summed to generate the Total WoE category (D). A final determination of appropriateness was based upon the following factors:

1. The WoE-D value equaled or exceeded a score of 7 (that is, as the statistical analysis in Table 4-2 shows, only those scoring above the sample mean were included, as a fairly high bar was desired to ensure appropriate study selection)
2. None of the other WoE score columns (A-C) values were scored as 1 (that is, all the WoE columns scored 2 or 3), essentially removing all studies with a Low score in any one of the WoE categories.

Table 4-2.

Descriptive Statistics of the WoE-D Column for Reviewed Studies

WoE-D Descriptive Statistics	
Mean	6.9449
Standard Error	0.2071
Median	8.0000
Mode	9.0000
Standard Deviation	2.3344
Sample Variance	5.4493
Kurtosis	1.5050
Skewness	(1.4918)
Range	8.0000
Minimum	1.0000
Maximum	9.0000
Sum	882.0000
Count	127.0000

Note: Descriptive statistics based upon the WoE-D scoring of all 127 studies examined for this review.

Factor 1 was based on the assumption that, if review totals were normally distributed, then only the upper range of the articles, at best, would achieve acceptance. As can be seen in Table 4-2, the sample set mean was under seven, meaning that those that were selected had achieved better-than-average total scores. Factor 2 eliminated any study that received a Low score for any of the WoE individual criteria, regardless of the WoE total score.

After these reviews and evaluations, 74 studies were assigned a “Yes” in the Accepted column; only these were included in the Research Question (RQ) analysis. All that were assigned a “No” response in that column were discarded. The entire selection process is summarized in the PRISMA Flow Diagram shown in Appendix D.

Data Collection Process

The data coding methodology employed in this review is somewhat unusual. While coding of each study is done, it is not in the standard form of, say, a thematic synthesis, where, as noted by Barnett-Page and Thomas (2009, p. 11), free codes of findings are organized into “descriptive” themes that are further interpreted into analytical themes. In the present study, the coding is based on a specific indicator, such as a learning algorithm identified within the study, and not an algorithmic theme. Thus, from an epistemological perspective, this study is similar in its approach to a quantitative analysis, using a *systematic* application of procedures to ensure objectivity and minimize bias, as contrasted with the *purposeful* approach of qualitative research (Bethel & Bernard, 2010, p. 235). Yet one can also view this process as taking a somewhat purposeful approach in that some of the naming conventions used for choice of algorithm are not standardized across studies. Thus, some judgment was used to determine where the specifications, even if named differently, represent the same use, and therefore categorizing them

collectively as one state. Moreover, even when the categories differ, the approach was similar for both the WoE coding and the RQ coding, as discussed below.

Still, this approach does not meet the definition of qualitative coding proposed by Miles, Huberman, and Saldaña (2014, p. 71): “Codes are labels that assign symbolic meaning to the descriptive or inferential information compiled during a study.” It might be more closely identified with Mays, Pope, and Popay’s (2005, p. 114) content analysis, a systematic technique for categorizing data which can be used in the synthesis of findings from multiple studies to count how often each category occurs in order to identify dominant findings and make generalizations. During the data-gathering discussion that follows, we will define how these categories were developed, with the intent, as with categorical coding, to supply “...unambiguous definitions that are consistently applied” (Oliver & Sutcliffe, 2012, p. 148).

Data Element Extraction for RQ Analysis

We now come to the RQ data collection, the process of compiling the information needed to answer this paper’s RQs. Eight data elements were collected from each of the selected studies, using the same NVivo 11 Pro for Windows coding tool that was used for the Weight of Evidence analysis. The data were assembled into a table for review (Appendix I, Research Question Analytics Database). The data items collected are as follows:

- What performance measure was used, and what was its outcome value? That is, what measure was used (e.g., prediction percent, ROC, sensitivity/specificity values, etc.) and what was the actual value attached to that performance measure (either a percent or a pair of percentages, the latter to signify both sensitivity and specificity values).

This information was typically extracted from the Results, Findings, or Discussion

- sections, and was evaluated to determine how best to categorize its assessment of ANN performance.
- Did the performance measure indicate a successful outcome? For the purposes of this study, three measures of success were applied:
 1. Full – the performance was fully successful as reported by the study author(s).
 2. Partial – the performance success was limited, due to mixed findings. For example, overall performance percent may have been strong, but sensitivity or specificity lagged, or the ANN’s overall performance was rated as effective but a competitive methodology (e.g., Logistic Regression or a Decision Tree) performed even better. Finally, if the reported measure did not meet the threshold established for this study (a performance rate of 80% - see the Data Discoveries section in Chapter 5 for detail on how that was determined).
 3. Failed – the performance was not successful; the results did not demonstrate that the ANN could classify the inputs to yield a reasonably predictive measure, as reported by the study author(s) – set at a performance rate of 60% (only slightly better than chance). Again, reference the Data Discoveries section of Chapter 5 for more detail.
 - Where did the ANN classifier perform well, and where did it not? Four data elements were targeted to identify the ANN application used.
 1. Study type – identified either diagnostic or prognostic.
 2. ANN application – identified the internal design methodology used for the ANN (refer to Appendix A for more detail).

3. Hybridization – identified when the ANN was used in conjunction with some other classification methodology (which would have augmented and, presumably, improved the classification ability of an ANN used as a stand-alone tool) and identified the augmenting tool.
 4. Algorithm use – identified instances when a particular ANN algorithm applied (internally) was expected to enhance the classification ability of the ANN processor (e.g., Levenberg-Marquart).
- The ANN study’s healthcare application – what was the disease, malady, issue, or clinical condition assessed within the study. This information also helped to determine if ANN success or failure was pertinent to particular healthcare issues, a part of RQ #2.

Data was collected by assessing the articles selected for review (using the same QSR International’s NVivo 11 Pro for Windows application that had been used for the WoE analysis), defining specific Nodes for each of the data items identified above, and transposing that coding into Appendix G – Research Question Analytics Database. The final analysis used the first three values to develop the answer to RQ #1 (how well did ANNs perform in the selected studies?) and used the last five values to address RQ #2 (where did they perform well or poorly?). That analysis is presented in Chapter 6 (Implications for Practice).

Examples of Data Extraction

In order to aid in transparency for this process, two studies were randomly selected from the 127 used for the RQ analysis, one that passed and one that did not (these two marked-up articles are reproduced in Appendix J). One includes only the Weight of Evidence (WoE) analysis, which it did not pass. The other appears twice, one for the Weight of Evidence (WoE)

analysis and the other (since it passed WoE) for RQ data collection. What follows is a narrative review of these studies as a means to exemplify the data extraction process. The format for Appendix J, for both studies presented, is that the related scoring matrix for each article shows at the top of the article's first page (including its final scoring) while the article itself follows. The referenced content is highlighted while, to the right, a coding bar makes clear the relation of that highlighted material to the scoring for which it applies. This format has been adapted directly from the NVivo application.

The first example in Appendix J is Andersson, Heijl, Bizios, and Bengtsson (2013). To begin, the highlighted part of paragraph two on p. 414 (Andersson et al., 2013) provides some explanation of the ANN design (WoE-C, Provide ANN Details). It notes that the ANN was developed for a previous study, hence one would need to refer to that cited study to obtain more information on how the ANN was designed and developed (e.g., what algorithm(s) were used, what software was used to create it, etc.). Next, the highlighted third paragraph on p. 414 of Andersson et al. (2013) specifically indicates that it is a diagnostic study, and that it compares clinicians and a “fully trained ANN” – therefore, both the first WoE-A question (Diagnostic or Prognostic Application) and the sole WoE-B question (ANN Related Study) could be answered directly from this paragraph. The final paragraph on that page (Andersson et al., 2013, p. 414-415) gives more indication of the ANN details (WoE-C) as well as a tidbit of information related to the ANN's validation process (WoE-A). Andersson et al. (2013, p. 415) provided ANN performance factors, addressing that issue for WoE-C. Finally, the number of patient cases was identified in the first paragraph of the Results section on that same page (Andersson et al., 2013, p. 415). That presents sufficient evidence to identify all of the WoE indicators, which allows us to make the acceptance determination. As noted in the matrix at the top, this work was rejected

from the present study since there was no validation process nor much in the way of specific ANN detail, even though some other assessments were acceptable for use within this study.

The second study (McLaren, Chen, Nie, & Su, 2009) includes a WoE analysis and an RQ analysis. Beginning with the WoE (second article in Appendix J), the first highlighted coding elements (McLaren et al., 2009, p. 2-3) note that the study assessed a diagnostic performance (WoE-A, Diagnostic or Prognostic Application) and compared the performance of ANNs against logistic regression (WoE-B, ANN Related Study). Next, McLaren et al. (2009) identified 71 individual patient cases selected as their sample (WoE-A, Number of Cases Examined) as well as information on the ANN design (WoE-C, Provides ANN Details), both on page three. Lastly, McLaren et al. (2009) described the ANN cross-validation used (WoE-A, Validation Process) on page four, and, finally, they provided outcome performance measures (WoE-C, ANN Performance) on page six. Thus, as can be seen in the matrix at the top of the article, this study met the WoE criteria for acceptance. (The only issue noted was the limited detail on the ANN structure, but what was provided was sufficient for this review).

As an example of RQ coding we revisit the McLaren et al. (2009) study, particularly the coding detail provided in the third article in Appendix J. While this is the same article coded for WoE review earlier, this pass looks at the detail differently. Again, with the scoring matrix at the top of the article's first page, and with the coding bars from NVivo to the right, one can identify where each highlighted section applies. While there is cross-over to the WoE analysis, the first highlighted section indicates that the coding here is different. In this instance (McLaren et al., 2009, p. 2), the highlight provides the answer to the question of what the clinical application on which this study focuses (the last part of RQ #2) – in this case, breast cancer. And just below that, another part of the RQ #2 Methodology question is addressed, whether this was a diagnostic

or prognostic study (this same node was coded for the WoE assessment to determine if it was *either* of those) – again, in this instance, it was a diagnostic study. One can proceed through the remainder of the article to examine the additional coding and to note how those reflect back to the scoring matrix located at the top.

Chapter 5 – Research Findings and Discussion

With the data collected, we commence with a discussion of the research findings. However, during the process of examining and evaluating the data, certain data variations were discovered that were difficult to reconcile using the methodologies chosen. Therefore, what follows first is a discussion on how those discoveries were reconciled with the methodology selected to ensure that the findings are consistent with the intent (examining ANN effectiveness).

Research Question Data Analysis

This now brings us to the primary discussion, that is, how the research findings addressed the two research questions (RQs) for this study. Each RQ is examined independently, followed by a discussion of the overall combined results from the two RQs.

Evaluation regarding research question #1 (RQ1).

The first research question posed in Chapter 1 seeks to determine how well ANNs performed in the research studies selected:

When ANN models have been used in healthcare studies, were they applied *effectively* as a high precision diagnostic or prognostic tool?

The first measure to be assessed in addressing this question is, how did ANNs perform in general, without regard to any other factors? Overall, 62 of the 74 studies (84%) were Fully successful, with the remaining 12 (16%) rated as Partially successful. While those measures alone can be interpreted as affirming ANN performance (in 84% of the studies examined, ANNs

achieved a predictive power of 80% or better), this still leaves too much room for error, given that this is a healthcare setting and the consequences of error are high, as detailed in Chapters 1 and 2. Thus, we need to dig further into the data to identify further evidence to support an assessment of relative effectiveness.

Analysis of diagnostic and prognostic study types provides a more nuanced evaluation, as shown in Table 5-1. For the 20 prognostic studies reviewed, the data indicated that 15 (75%) were Fully successful, and for the 54 diagnostic studies, 47 (87%) achieved Fully successful; as with the aggregate above, these are fairly high marks. Some analysis of the differences between those types of studies will be examined in greater detail later in this chapter, but at worst, 75% of the time ANN tools in research appear to achieve the 80% performance rate level or better. Nonetheless, further analysis still seems warranted.

Table 5-1.

ANN Type Category Performance Within the Analyzed Studies

Count of ANNs by Type	Success				Totals	
	Partially		Fully		Count	Percent
ANN Type	Count	Percent	Count	Percent	Count	Percent
Diagnostic	7	13%	47	87%	54	73%
Prognostic	5	25%	15	75%	20	27%
Grand Total	12	16%	62	84%	74	100%

Note: Categorized as Diagnostic or Prognostic study. Partially: <80% predictive power; Fully: ≥80% predictive power.

Taking another perspective, when all of the studies are examined by their reported predictive power (for Sens-Spec dual measures an average is taken as a way to aggregate to a single measure) the degree of success seems to be even stronger. The mean predictive power across all 74 studies (89.33%) is shown at the end of the Actual Predictive Power by Study table in Appendix L. Thus, regardless of level of success, the studies showed an average of an almost 90% predictive power, a more convincing measure than those previously discussed. Analyzing

the results by the measurement tool used, as shown in Table 5-2, indicates that a high performance appears across that dimension as well. Within those studies classified as Fully successful, an average predictive power of 91.72% was achieved.

Table 5-2

Average Predictive Power by Performance Measures

Average of Predictive Power Measurement	Successful	
	Fully ($\geq 80\%$)	Partially ($< 80\%$)
AUROC	91.37%	74.40%
Prediction %	92.82%	76.58%
Sens-Spec	90.51%	83.75% ^a
Spearman	84.00%	—
Mean	91.72%	74.03%

Note: Average predictive power found in the 74 studies examined, categorized into Fully (80% or higher) or Partially (under 80%) successful.

^aThis discrepancy in predictive power scoring is due to Study IE #172, as noted earlier under the “Unanticipated Data Findings from the Reviewed Studies” section, specifically referencing Footnote 8.

Hence, with an overall predictive power of roughly 90%, and for those ANNs deemed Fully successful, a predictive power exceeding 90%, the answer to RQ1 is that these results strongly suggest that ANNs performed effectively as a high-precision tool in research studies. And for those in the Partially successful category there was still a predictive power well above what could only be attributed to chance, as indicated by the mean predictive powers noted in Table 5-2 (the lowest being 74.4%). Of additional significance is that no cases among the entire sample were determined to be unsuccessful (that is, none had a predictive power of less than 60%). This latter observation may be due to publication bias or the peer-review process.

Since some of the studies selected for this review included comparisons between predictive approaches (e.g., ANNs versus Logistic Regression), it might be expected that, whether or not the ANN tool was the best performer, there could have been actual failures of

ANN tools (that is, achieving less than 60% predictive power) in some of those comparative studies. However, none of the 21 studies that undertook comparisons with alternative methodologies reported an ANN success rate below 60%, and 16 of the 21 (76%) reported the ANN tool as Fully successful, as shown in Table 5-3.

Table 5-3.

ANN-specific Performance Measures in Comparative Analysis Studies

Comparison Studies	Success		Total
	Fully	Partial	
MLPBPNN	11	3	14
MatLab	3	1	4
SPSS	1	—	1
WEKA	1	—	1
EasyNN	—	1	1
Total	16	5	21
	76%	24%	100%

Note: ANN-specific performance measures in studies that undertook comparative analysis between ANN-based methodologies and others (alternate machine-learning systems, statistical analyses, or humans). Full: $\geq 80\%$ predictive power; Partial: $< 80\%$ predictive power.

Further, of the 21 comparison studies, only two presented cases where the ANN did not perform at least as well as the alternative(s), as seen in Table 5-4, and in both of those studies, the ANN was determined to be Fully successful. Thus, in 19 of the 21 studies (about 90.5%), the ANN tool equaled or exceeded the alternative. This was true even for studies where the ANN's predictive power was at the Partial level.

Table 5-4.

Relative ANN-tool Success in Comparative Studies

Comparison Studies	Success		
	Full	Partial	Total
Better	12	2	14
Equal	2	3	5
Worse	2	—	2
Total	16	5	21

Note: Relative success of ANN tools in studies that undertook comparative analysis between ANN-based methodologies and others (alternate machine-learning systems, statistical analyses, or humans). Full: $\geq 80\%$ predictive power; Partial: $< 80\%$ predictive power

It can therefore be concluded that ANN success is quite strongly supported within the evidence collected for this review.

Evaluation regarding research question #2 (RQ2).

The second RQ in this study highlights contributing aspects of both successes and failures of ANN tools in research studies:

Of those ANN studies analyzed, under what conditions and/or what applications have they tended to perform with greater effectiveness (and, conversely, where have they not done so)?

As was clear from the analysis provided for RQ1 above, there was an absence of ANN tool failures (with failure defined as not performing *better* than a predictive power of 60%, that is, a little better than random). First, we examine the study type and its influence on ANN performance, followed by other attributes captured during the data-gathering process.

Using the study type classification as a subgrouping (Table 5-1), it was found that while most of the studies examined (54 of 74, or 73%) were of a diagnostic type, a substantial number

(20, or 27%) were prognostic. To determine if these ANN type subsamples were connected in some way that relates to performance, a chi-square (χ^2) independence of categorical values test (Burns & Burns, 2012, pp. 334-340) was employed. The null hypothesis in this test is that the categorical variables of study type (diagnostic/prognostic) and success (full/partial) are independent. The χ^2 test for independence calculation shown in a contingency table (Table 5-5) indicates that these results are not significant at the $p < 0.05$ level, suggesting that these variables cannot be shown to be dependent (we cannot reject the null hypothesis), limiting our ability to ascertain if either study type indicates a stronger fitness for ANN performance. While diagnostic studies seem to perform better at the Fully successful level percentage-wise, there is insufficient evidence to establish that as a significant relationship. This may be due, at least in part, to the small sample size for prognostic studies.

Table 5-5.

Chi-square Contingency Table for Performance/Type Independence

	Diag	Prog	Total
Partial	7	5	12
(Exp)	(8.76)	(3.24)	
[Chi2]	[0.35]	[0.95]	
Full	47	15	62
(Exp)	(45.24)	(16.76)	
[Chi2]	[0.95]	[0.18]	
	54	20	74

$$\chi^2 = 1.5564$$

$$p = 0.212193 \quad [\text{cannot reject at } p < 0.05]$$

Note: Contingency table illustrating a chi-square (χ^2) calculation to determine the independence of the performance measures from the ANN study type.

Table 5-6 shows the full complement of ANN tools employed within the 74 reviewed studies, and the frequency with which each was used across the entire study group, revealing a

high concentration of studies that used just a few of the ANN tools listed. MLPBPNNs dominated with 41 (55%), which, as was noted at the outset, is of concern since that is a generic category for unspecified tools (albeit, those that simply met the MLPBPNN design criteria). Within that category, 34 of the 41 studies (about 83%) achieved Full success level, which is nearly identical to the performance across the entire collection (62 of the 74 studies, or about 84%). The next five tools listed (MatLab, StatSoft, SPSS, Neurosol, and WEKA) represent an additional 26 uses, about 36% of the studies. Within those 26 studies, 92% performed at a Full success level, almost 10% higher than the full dataset; only two (one each for MatLab and Neurosol) fell into the Partial category, representing only about 8% of these cases.

Table 5-6.

ANN Application Software (Tools) Used in the Clinical Study Sample

Count by ANN Tool Tool Name	Success			Percent
	Partial	Full	Total	
MLPBPNN	7	34	41	55%
MatLab	1	15	16	22%
StatSoft	—	4	4	5%
SPSS	—	2	2	3%
Neurosol	1	1	2	3%
WEKA	—	2	2	3%
C++	—	1	1	1%
Math Works	—	1	1	1%
EasyNN	1	—	1	1%
CU-ANN	—	1	1	1%
Neurointell	—	1	1	1%
ANNES	1	—	1	1%
neURON++	1	—	1	1%
Grand Total	12	62	74	100%

Note: Includes categorization of that performance measure into Partial or Full success. Partial: <80% predictive power; Full: ≥80% predictive power.

The difference in performance between the MLPBPNN group (83% Full success) and this second group of five (92% Full success) warrants additional examination to see if any patterns emerge.

Within that group of five tools, the numbers for all but the MatLab tool are quite small, so it would likely be difficult to generalize those findings (noting that the sample sizes were too small to attempt any meaningful statistical analysis such as the χ^2 test done earlier). Indeed, Neurosol by itself was used in only two studies, and one of those was a Partial success. The other three tools achieved a Full success level in all cases. Of the 16 MatLab studies in this group, all but one were a Full success, suggesting some potential for that tool, but again, that sample is too small to allow for generalization. Thus, while the uniquely identified tools performed well, further research is required to determine if that success is of significance to ANN tool usage generally.

Next, we evaluated the clinical applications specified in the studies under review. As noted earlier, the study set represents a broad variety of clinical applications (see Appendix K). However, 24 of those studies relate to some form of cancer, the largest subgroup identified, which may warrant further examination of this dataset sub-classification. These cancer-related applications are presented in Table 5-7 and graphically in Figure 6.

Table 5-7.

Performance of Cancer-related Clinical Applications

Clinical Application Performance				
Study Application	Success			Percent
	Partial	Fully	Total	
Breast Cancer	1	10	11	14.86%
Metastatic Cancer	1	2	3	4.05%
Lung Cancer	1	1	2	2.70%
Liver Cancer	—	2	2	2.70%
Brain Cancer	—	2	2	2.70%
Basal Cell Carcinoma	—	1	1	1.35%
Oral Cancer	—	1	1	1.35%
Colorectal Cancer	—	1	1	1.35%
Esophageal Cancer	—	1	1	1.35%
ALL OTHERS	9	41	50	67.57%
Grand Total	12	62	74	100.00%

Note: Clinical applications relating to some form of cancer, with their levels of success, with all other clinical applications grouped into one category (ALL OTHERS). Partial: <80% predictive power; Full: ≥80% predictive power.

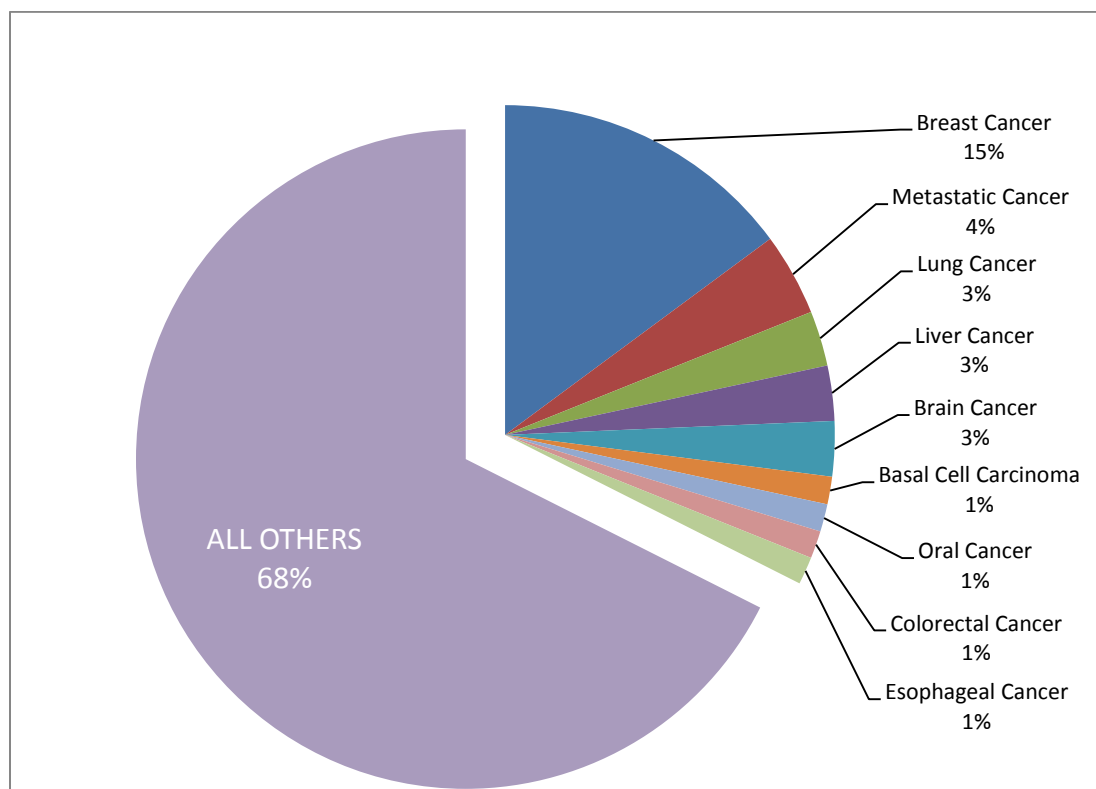


Figure 6: Clinical applications relating to cancer within the reviewed studies, indicating their frequency of occurrence within the study set.

As can be seen in Table 5-7 and Figure 6, the cancer studies as a group represent almost one third of the full study dataset. Because this is such a large portion of the dataset, investigating the predictive power for this group differs from that of the entire sample (again noting that these individual sample sizes are too small for detailed statistical analysis). Review of this data shows that 21 of the 24 cancer studies achieved a Full level of success, yielding a rate of 87.5%, slightly better than the 84% success rate of the entire dataset (as indicated earlier in Table 5-1). Within this cancer-related sub-group, however, 21 of these studies (91.3%) were at the Full success level, and only two (8.7%) were at the Partial level (Table 5-8).

Table 5-8.

Performance of Cancer-related Diagnostic Clinical Applications

Diagnostic Clinical Applications Cancer Subset	Success			% of All Studies
	Full	Partial	Total	
Breast Cancer	10	—	10	13.51%
Brain Cancer	2	—	2	2.70%
Metastatic Cancer	2	—	2	2.70%
Liver Cancer	2	—	2	2.70%
Melanoma	1	1	2	2.70%
Lung Cancer	1	1	2	2.70%
Basal Cell Carcinoma	1	—	1	1.35%
Colorectal Cancer	1	—	1	1.35%
Oral Cancer	1	—	1	1.35%
Total	21	2	23	31.06%
	91.3%	8.7%	100.0%	

Note: Clinical applications identified as relating to some form of cancer, but limited to only those determined to have a diagnostic focus, and listed by their descending level of success. Full: $\geq 80\%$ predictive power; Partial: $< 80\%$ predictive power

With a Fully successful rate 7.3 percentage points higher than the full dataset, these diagnostic cancer studies seem to represent a particularly strong category for ANN success. Thus, further research focused on diagnostic cancer studies and ANN performance would seem to be of value. What is more, almost half the studies in this diagnostic cancer group are specific

to breast cancer, and all of those were Fully successful. Here, too, the sample size (only 10) is too small to support in-depth statistical review, but the data at least suggests the need for additional, highly-focused research into diagnostic breast cancer studies.

Finally, the last remaining step in the RQ2 analysis was to identify which algorithms and which hybrid processes proved to be better performers. Only a small minority of the studies provided sufficient detail about the ANN tools to allow identification of the neural network algorithm or hybrid processes used (16 and 8 cases, respectively, as shown in Table 5-9 and Table 5-10). Since these features were identified in so few cases, little can be done to determine whether the algorithm selected, or any hybrid process, was effective in augmenting the performance of the ANN tools in these studies. Nor is there sufficient data to establish that further detailed research in those areas is warranted (other than to highlight the need in future research to identify the measures used).

Table 5-9.

Algorithms Used in Reviewed Studies.

Algorithm Usage	Success		Total
	Partial	Full	
Leven-Marq	—	5	5
GD	—	2	2
Adapt Prop	2	—	2
Genetic	—	2	2
CG	—	2	2
Markov Chain	—	1	1
Langevin	—	1	1
Bayesian	—	1	1
Total	2	14	16

Percentage of All Studies: **22%**

Note: Algorithms identified in the studies reviewed, including their level of success. Partial: <80% predictive power; Full: ≥80% predictive power.

Table 5-10.

Hybrid Processes Applied in Reviewed Studies.

Hybrid Processes	Success		Total
	Partial	Full	
ICA	—	2	2
LZ	—	1	1
MPANN	—	1	1
PSO	—	2	2
Wavelet	—	1	1
ELM	—	1	1
Expert	1		1
Total	1	8	9

Percentage of All Studies: **12%**

Note: Hybrid processes identified in the studies reviewed, including their level of success. Partial: <80% predictive power; Full: ≥80% predictive power.

Unanticipated Data Findings from the Reviewed Studies

As discussed in Chapter 2, the measure of ANN effectiveness was generally left to the domain of the research study author(s), and if the peer-reviewed study identified the outcome as one that had a predictive (i.e., effective) impact, then that was to be accepted for purposes of this analysis. Still, that leaves open varying *gradations* of effectiveness among the studies. As posed in Chapter 4, it was noted that there was no apparent, meaningful way to differentiate relatively close predictive rates (e.g., 88.5% versus 84.3%). Moreover, after the studies were examined and the data collected, the analysis did not help to definitively clarify that issue. However, a preponderance of the studies (62 of the 74 studies reviewed, or roughly 84%) noted a performance measure (hereafter referred to as the *predictive power*) that was at least 80%. Hence, it was determined that this study would use a threshold of 80% as a delimiter to differentiate “Fully” successful studies (predictive power of 80% or higher) and “Partially”

successful studies (predictive power exceeding 60% but not reaching the 80% threshold).⁷ Of note, many of the studies analyzed for this review used a similar conceptual, if not explicitly stated, threshold to determine if ANN performed effectively. Thus, using this classification allowed for the differential performance analysis presented later in this chapter.

Still in the context of performance, another aspect came to light when examining the various predictive power measurement methods encountered, which are presented in Table 5-11. Only four measurement methods were used across all 74 studies, and two of those constituted the great majority (89%), Prediction Percent and AUROC. However, the third measure on the list – Sens-Spec (Sensitivity-Specificity) – is a dual scored measure, which creates a problem when doing comparisons. The first measure (Sensitivity, or true positive rate) represents the percentage of the ANN's positive outcomes that were for actually positive cases, while the other (Specificity, or true negative rate) represents the percent of ANN negative outcomes that were ultimately determined to be negative.

⁷ Note that those studies reporting predictive power below 60% were not considered to have achieved a performance that significantly exceeded random (50%), and, fortunately, no such studies were encountered (which we suspect was due to the peer-review process and/or publication bias).

Table 5-11.

Number of Occurrences of Each ANN Performance Measure Found

Performance Measure	Success			% of Total
	Partially	Fully	Total	
Prediction %	6	30	36	49%
AUROC	5	24	29	39%
Sens-Spec	1	7	8	11%
Spearman	—	1	1	1%
Total	12	62	74	100%

Note: Includes categorization of that performance measure into Fully or Partially successful. (Partially: <80% predictive power; Fully: ≥80% predictive power).

As a way to deal with these two measures for the purpose of categorization (that is, Fully or Partially successful), each measure was evaluated independently. Thus, it was decided that when *either* of the Sens-Spec values fell below the 80% threshold then that measure was considered *Partially* successful. Only eight cases used a Sens-Spec measurement (about 11% of the aggregate) and each was evaluated using this conservative approach.⁸

The remaining method, a Spearman's rho (or Spearman's Rank Order Correlation), was encountered only once. In that study, the ANN achieved an overall Spearman's rho (r_s) of 0.84 ($p < 0.0001$, $n = 100$), which would suggest this result has a very large effect size at the 0.05 level of significance (Burns & Burns, 2012, p. 358). The challenge this particular study presents, however, is that the comparison used was to that of a panel of expert clinicians and not to the actual outcome of the patient's true diagnosis. Therefore, the outcome measure for this study

⁸ Note that seven of the eight Sens-Spec measures met the 80% threshold and were, therefore, classified as Fully successful, while the one Partial was due to *one* of the two measures being below 80% (reference Study ID #172 in Appendix I).

can only be as strong as a skilled human interpreter, which may reduce its value as a performance measure. For the purpose of the current paper, given that this particular study achieved a performance level at least as effective as a human expert, and that its predictive power exceeded the 80% threshold, the study was categorized as Fully successful.

Another challenge to the analysis was identified when examining the ANN tools (software) that were used (Table 5-6, shown earlier). Over half (55%) of the studies did not explicitly specify the software application, even though in most cases the tool description strongly suggested that of a generic Multilayer Perceptron Back-propagation Neural Network (MLPBPNN), as described in Appendix B. Of note, the MLPBPNN designation indicates only that a common ANN design was applied to that particular study, while in reality it could have been any of the other tools listed in Table 5-6, or even a tool that is not on that list. Therefore, the MLPBPNN category was created to represent the non-specific “tool” that fit MLPBPNN design parameters, which resulted in 13 tool categories shown in Table 5-6. Thus, for over half of the studies reviewed, the lack of specificity made it difficult to determine comparative performance measures.

Finally, among the 74 studies reviewed, researchers examined a broad set of diseases or maladies (Appendix K), what we refer to in this paper as the ANN’s *clinical application*. A total of 54 unique applications were identified, with Breast Cancer being the most common (11 of the 54, or about 20%); only eight others were represented more than once. While the breadth of applications may seem to support the broad use of ANN tools in clinical research, this finding makes it more difficult to undertake a comparative examination to determine which application(s) performed better. The evidence was spread too thin to provide direct support for

any particular clinical application's use of ANNs in practice. A particular review in the trade literature identified three global problems with the use, in general, of AI tools in practice.

Discussion

This brings us to the full assessment and synthesis of the systematic review process and to a determination of the answers to the two research questions. In the case of RQ1, it seems clear from the data that ANN tools are generally effective and perform well across a spectrum of clinical applications. Across the 74 studies reviewed, 84% attained a Fully successful level with an average predictive power of 91.5% (or, looked at from a different perspective, an average error rate of 8.5%). As noted in Chapter 3, the diagnostic error rate in clinical medicine is approximately 15% (Berner & Graber, 2008, p. S3). Thus, for the Fully successful ANN tools identified in this review, their average error rate was about half that of the observed error rate in diagnostic clinical medicine. This suggests that these results can be used in support of the proposition that ANNs can function as high-precision tools across a wide variety of clinical applications as a means to augment human diagnostic and prognostic assessment.

Regarding the more complex question, RQ2, attempting to determine where ANNs performed better or worse was somewhat problematic, in part due to the broad spectrum of their use within the study set. While this study initially identified 364 studies, the conservative criteria for inclusion reduced that number to 74, and when these were evaluated across the dimensions of study type, ANN tool, application, and even ANN performance measure, there was not sufficient data to determine any clear relationships. This was particularly an issue for those measures that were not well described in the studies (e.g., in studies that did not indicate specific ANN algorithm or hybrid application selections). Thus, one could suggest that the scope of the present ANN study was somewhat undermined by the broad success of ANN usage in

clinical research, as it is difficult to assess performance across a very wide and diverse range of applications, and even more difficult when key data were not included in the analyzed studies.

Nonetheless, the data yielded some strong suggestions that certain opportunities exist for legitimate exploration of ANNs, such as their use in diagnostic cancer studies, most specifically in the case of breast cancer diagnosis. The analysis further seemed to suggest that ANN performance may be stronger in diagnostic rather than prognostic applications, and that specific software tools that were used most frequently also had a high success rate, but that there were insufficient cases to test such assertions. Thus, although this study could not resolve several questions to the optimal level of detail, the results clearly suggest opportunities for future targeted reviews.

Chapter 6 – Implications for Clinical Practice and Healthcare Management

Having shared the findings of this review, and provided some discussion of the detail as well as the limitations of those findings, we now turn back to the topic of greatest interest to clinicians and healthcare managers – the implications of these findings to practitioners.

Implications of the Research Question Findings

The purpose of this systematic review was to examine ANN classification tools applied to diagnostic and prognostic assessments in the research literature and to determine the potential for their use in clinical practice. The objective was to determine “what works” from a practice perspective – specifically, were the ANN tools demonstrated to be effective (RQ #1), and where was their performance better or worse (RQ #2). While ANN tool use within the research literature suggests that they are successful, there is a dearth of evidence regarding their application or their efficacy in clinical practice (Chapter 2). The findings of this review, as noted in the RQ #1 discussion (Chapter 5) strongly suggest that ANNs are, indeed, effective tools,

while the RQ #2 discussion (also in Chapter 5) provides evidence as to which clinical applications might be better initial candidates for practice (specifically, diagnostic breast cancer assessments). Overall, given the findings herein, the evidence seems to give solid support to the use of ANN tools by clinical practitioners in diagnostic and prognostic assessments.

Lack of Clinician Familiarity with ANN Tools and Concepts

However, what does this mean to the practitioner and clinical managers, and what does their lack of familiarity with the topic of ANN research (as noted in Chapters 1 and 2) suggest is needed? Several peers with whom I have engaged in primary healthcare-related studies have corroborated the lack of ANN tools in practice. As well, in three presentations of studies for which I have been a co-investigator (Belliveau, Axt, & Seetharama, 2015; Belliveau, Seetharama, & Axt, 2015a; Belliveau, Seetharama, & Axt, 2015b), I did not have a single encounter with anyone who had knowledge of ANN use in practice. Further, none of those presentations yielded contacts with any clinicians who had experience with or who had employed ANN technology, and most were completely unfamiliar with the concept of ANNs generally. Thus, it is an uphill climb not only to educate clinicians and managers about the capabilities of ANN tools, but also to engender their trust in its application, especially in a field as sensitive to error as healthcare. This harkens back to the discussion of barriers to change in Chapter 2.

First, as discussed in Chapter 1, a generally accepted reason for the limited use of ANN technology in practice has to do with the “black box” assertion from which ANN implementations suffer. Since an ANN develops its network through an iterative learning process not readily apparent to its user, and with each training dataset often being unique to that study, the ANN’s learning mechanism is not transparent even if it is later determined, through

outcome evidence, to be effective. Indeed, the ANN algorithm's iterative calculations are exceedingly difficult to visualize because unlike, for example, an analog thermometer, the *manner* by which the results are arrived remain unknown to the user (Gant et al., 2001, p. 345). Thus, while the ANN process can be generally described, it differs from a formulaic algorithm such as that used for logistic regression. There is no method to determine a clear-cut path of logistic reasoning that would make sense in the ANN model. Therefore, the ANN appears to observers as a "black box" which inexplicably arrives at its learned conclusions, making it seem to these observers as untrustworthy for scientific use, especially where human lives are concerned. A great many of the ANN studies cited in this systematic review addressed this concern as part of their methodology discussion; that is, that the "black box" issue was of concern to those who employed ANNs in their work, even though ANNs' efficacy within the individual studies had been demonstrated.

Piloting Approach and Practice Development Center

In order to address this lack of familiarity and trust it would be valuable for pilot implementations to be engaged by willing clinical practices. Given the performance metrics noted in this review, we expect to be engaging other clinicians in ANN piloting, particularly those for whom this study has suggested good potential for success. This would include those physician practices involved in diagnostic breast cancer screening, as that specific practice model performed particularly well within the present review (Chapter 5). That would begin to address this resistance to change, noted back in Chapter 2, in the manner suggested by Lewin (1947) in terms of systemic change through providing evidence of success. In addition, it is expected that such piloting activity, when brought before the medical community through published reports and presentations, would begin to alter the teleological resistance and concern that practitioners

have expressed about this technology (sociocultural acceptance through common usage – Burnes, 2004; Van de Ven & Poole, 1995). Bringing clinicians directly into the pilot allows for the integration of social, psychological, and organizational implications of practice change as part of the implementation process, identifying and resolving issues such as workflow considerations under controlled (that is, pilot) conditions.

As was also discussed in Chapter 2, ANN technologies are disruptive innovations, and as Christenson's (1997) work suggests, a separate organizational pilot model presents the best means of practice development and, ultimately, acceptance. This requires managerial agreement and investment in this technology in order to transform it from a theoretical proposition to an evidence-based practice model. Indeed, as a follow-on to this dissertation, a co-investigator on a previous study and I are working to establish a center of rehabilitative medical practice design at our current workplace, Hospital for Special Care (HSC). This center is expected to engage informatics professionals, seasoned research statisticians, and HSC clinical staff in order to build specific practice models that employ ANN tools in the manner shown by the DRAWN model described in Chapter 3. We have already petitioned HSC, which has a significant population of patients with rehabilitative needs, to assist and support us in that effort, and we have also engaged a local university statistics professor and several hospital-affiliated physicians. This center's mission would be for these skilled individuals, working as a team, to obtain and/or develop ANN tools based upon historical clinical datasets and, most significantly, to pilot those tools within the physician's clinical practice at the institution. So far, at this early stage, the proposal has been well received by all of the HSC executives involved, and they are currently investigating funding options for us. In the longer term it is hoped that this will eventually become a center of excellence for ANNs in clinical use, demonstrating piloted applications the

DRAWN model's use in rehabilitation, and eventually in other clinical arenas beyond that practice (such as diagnostic cancer screening).

Existing Research Demonstrating Practitioner ANN-use Success

Next, several of the studies analyzed for this paper demonstrated the applicability of ANNs to practice. One study in particular (Olsson, Ohlsson, Ohlin, Dzaferagic, Nilsson, Sandkull, & Edenbrandt, 2006) employed an ANN (MLPBNN) in evaluating specific signals registered by a 12-lead electrocardiogram (ECG) to assess acute coronary syndromes in urgent and emergent care settings. This study used two highly experienced cardiologists to represent the “gold standard” of diagnostic assessment, and compared their performance to that of the trained ANN and to the assessments of physician interns who had less than one-year of clinical experience. The study's purpose, as the authors stated, “...was twofold, to develop an automated tool for the analysis of ECG...[and] secondly, to assess how the tool could influence the interpretation of ECGs by physicians [i.e., interns]” (Olsson et al., 2006, pp. 151-152). The performance statistics provided in the study were very telling (Olsson et al., 2006, pp. 151-154): The ANN performance reached a 98% overall prediction rate, with 95% sensitivity and 88% specificity, and with a good correlation to the experienced cardiologists' assessments. The interns, however, on average reached a 68% sensitivity and a 92% specificity performance, misclassifying a fair number of the diseased patient cases. A corollary part of the study, giving the interns access to the ANN as part of their assessment, improved their sensitivity to 93% (a 37% increase⁹) with only a small decrease of their specificity to 87%¹⁰ (very close to that of the experienced cardiologists). This provides a clear demonstration of the ANN's direct application

⁹ From a sensitivity of 68% increased by 37% to 93% ($0.68 * 1.37 \approx 0.93$)

¹⁰ From a specificity of 98% decreased by 5% down to 87% ($0.98 / 1.05 \approx 0.87$)

within practice to improve a clinicians' overall performance in diagnostic classification. Thus, the Olsson study provides one established evidentiary use case for clinicians and managers unfamiliar or ill at ease with ANN tools.

Post-study Implications for ANNs in Practice

The present study gives both the evidence necessary to support ANN performance and application as well as a requisite model for practice. The ANN clinical decision-making process integration is modelled on the DRAWN model (enhanced “Data Refinery” originally developed by Gant et al., 2001) as described in Chapter 3 and illustrated in Figures 3 and 3a. How that model can be integrated into practice is the topic of the next discussion, adopting the “new way of thinking” posited by Reio (2009).

ANN Application Delivery

As noted in Chapter 3, for several of the most common research software applications used, the developed ANN can be employed through a run-time application (the network exported through distributable software that can take the required inputs and produce the ANN's output within an application window, much like a calculator). Thus, the development effort for many diagnostic or prognostic ANNs has already been accomplished because researchers, such as those identified within this paper's systematic review studies, built and proved them as part of their study. This limits the cost of using that particular ANN to whatever those researcher(s) determine as amenable to their own research practice. The ANN could be delivered as a licensed package that can be imported and applied by a run-time version of the software under which it was developed, noting that run-time versions of software are usually available at a minimal cost as compared to the full development application. This approach is very similar to those employed by mobile app vendors. Thus, the distribution of ANN technology from research to

practice can follow an established model that has been applied to commonly used technological applications (e.g., the mobile phone or desktop app).

Implementing the DRAWN Model

Under the DRAWN model as described in Chapter 3, the ANN can be understood as a consultative agent to the diagnostic or prognostic assessment process, off-setting the influence of human bias mediators (such as attempts at satisficing and sensemaking, or a heuristic thinking process). The decision-making bias is counterbalanced by the evidenced-based outcome provided by the ANN, as described in the clinical decision logic chain presented in Chapter 3 (and Appendix E). The ANN component of this model can be implemented through several different approaches, each progressively more tightly tied to an associated clinical system:

1. As a stand-alone application where the clinician inputs parameters resulting in an output decision from the ANN.
2. As an integrated application to an Electronic Medical Record (EMR) system that sources its inputs from the EMR directly upon request of the clinician.
3. As an embedded EMR application that triggers the ANN network to “fire” at a pre-determined time based on the clinical workflow.

The advantage of the first option (stand-alone application) is that it requires the least amount of integration and thus has the greatest flexibility. Using this approach, the clinician would intentionally invoke the preset ANN at his or her own discretion, and would be prompted for (and manually enter) the necessary ANN parameters from the case being examined (the case of interest). The ANN would produce a particular result, a categorical selection, and display that to the clinician for review. Since there is only minimal clinician involvement, and presumably no patient-identifiable information required (which would require a HIPAA-compliant interface),

this tool can be employed alongside an existing EMR system or as a stand-alone tool in a clinical practice. Indeed, such an application need not be co-resident with the clinician. It can be developed as a client-based system (requiring clinicians to load software on PCs or mobile devices) or as a Web application.

The major flaw in this approach is its introduction of human entry error into the process. A typographical or transposition error can alter the ANN outcome greatly, with no mechanism to provide the clinician with any related warning – the ANN simply responds using the values presented. The only way to mitigate such entry errors is to build some logic into the data entry tool that validates that entry, where possible – for example, ANN parameters that have a standard range of values can have their entry limited to that range. Yet that is still an imperfect solution since, for example, the entered value may be appropriate for certain cases (adult patients) while highly inappropriate, and indeed, potentially fatal, in others (pediatric patients).

The next option (integrated application to an EMR system) would be similar in presentation to the first, however, the parameters would, if available, be pre-populated with case-specific content from the EMR system itself. This would require building an interface to the EMR to extract the necessary data fields. In similar clinical applications this has been configured as an Application Program Interface (API) or, especially in the healthcare setting, a Health-Level 7 (Application Level) interface, and the same would be expected here. Typically, these would be “pull” applications – that is, they would draw (pull) information from the source system (EMR) but not return data back, as that would require a more complex level of interaction between the run-time ANN and the EMR. The ANN application would still be under the control of the clinician (e.g., the clinician would click on an activation button to trigger the

ANN) though it could also be automatically triggered by other clinical events, depending on how the workflow were designed at the time of implementation.

This may somewhat mitigate the entry error issue noted in the first option as EMR applications often self-audit discrete values. For example, valid test value ranges are typically used to validate or audit test results within the EMR, such as for common laboratory tests. On a more complex level, data analytics have been implemented in some more robust EMR systems that interactively flag unexpected results or outcomes, such as suggesting a diabetes diagnosis outcome when their A1C value is 5.1, well within normal range, which would predispose the ANN inputs to auditing and validation before being brought into use. The Clinical Document Improvement (CDI) Engage application from M*Modal (2017) is one such EMR real-time auditing add-on.

The third option (an embedded EMR application) is the most complex to implement, but renders the ANN as fully integrated within the EMR platform and thus potentially invisible to the clinician, except when it presents the outcome for the case of interest. Not only would the source data be auto-populated into the run-time ANN, but the results would be fed back into the EMR and retained as part of the clinical record. While this may be an optimal solution, its adoption is much more problematic. In essence, the clinician's control point (deciding if and when to employ the ANN) would be overridden by the EMR's workflow, and (as suggested in the literature) this is not likely to be well received by medical professionals. Indeed, the intent of the DRAWN model is to ensure that the clinician retains decision-making control. It does not force them into a less holistic decision based solely upon the ANN's historical data from its training cases (which might not fully represent the context for this particular patient case). Still,

the workflow could still require a final, authoritative, decision by the human clinician, even in this case.

Regardless of the implementation methodology, the DRAWN model can have a broad reach, since run-time ANN tools typically require only that the application be made available to its user. The run-time ANN tool is easily adapted to Web technologies so that the ANN itself does not need to be co-resident with its user or with any associated EMR application. Thus, a hospital or clinic in any part of the globe can access this technology on a mobile phone with little more than a link to the Internet. This is a clinical consultant model that has only been dreamed of previously; one that does not require human involvement beyond the requesting clinician, yet that provides evidence-based consultation while maintaining the human clinician as the final decision maker.

The Evidence in Support of ANN Value to Practice

As questioned in Chapter 3, why invest in ANN technology if, at best, it does nothing more than replicate the knowledge, expertise, and experience of highly-skilled clinicians? First, the ANN is evidence-based, so it is not subject to bias as humans are; second, there may be information hidden in the patient's data that the clinician was unaware of, and that influenced the ultimate decision (e.g., some unrecognized relationship between the outcome and an input measure). Within the DRAWN model, the ANN functions as an "outsider" (Bazerman & Moore, 2013), providing a perspective that may be different from the clinician's point of view. As such, one can view the ANN as an error-detection aide applied to clinical decisions. As with any quality improvement approach, there is an expected performance gain due to the ANN assessment's influence, hence overall patient outcomes should improve as well.

DRAWN Practice Implications

However, the benefits derived by ANNs, as suggested here, go beyond clinical performance measures. As discussed in Chapter 1, there are financial implications of incorrect diagnostic or prognostic assessment, such as an increase in the number of healthcare services required as well as their associated costs, both direct and indirect. The risk-avoidance benefit, as evidenced by the ANN performance vis-à-vis the interns in the study by Olsson et al. (2006) referenced above, in preventing diagnostic misclassification, a performance improvement of 37% in those cases where disease was actually present. Those misclassifications would undoubtedly have incurred follow-up patient visits for urgent or emergent care as patients' conditions worsened or were exacerbated by lack of medical attention, thereby inflating the cost of care well beyond what it would have been had the correct diagnosis made at the outset. Consider the following chest-pain scenario:

A patient was discharged from the Emergency Department (ED) to home, whereupon the cardiac symptoms that prompted the ED visit initially had returned. The relapse might require an EMS call, an additional ED assessment, and further laboratory and cardiac tests, as well as the time and attention of a number of clinicians (interns, nurses, ED operational staff, ancillary staff [laboratory, radiology, and cardiology], and administrative staff). All of this extra activity, much of it replicative, inflates the overall healthcare cost for that patient beyond the initial visit.

While a cost analysis for such a scenario was not found in the literature (likely due to great cost variability based on service/practice), it is reasonable to suggest that in the absence of ANN applications, significantly higher direct costs are likely in healthcare delivery systems. As noted in Chapters 2 and 3, the literature on quality assessments of clinical practice corroborates this scenario quite well. This, of course, does not include the legal and reputational risks associated with a missed diagnosis. Thus, as noted in Chapter 3, if the DRAWN model

implementation can be shown to reduce risk, loss, and error, and improve patient outcomes, while not incurring excessive costs (as measured by the above benefit), then it is certainly a justifiable enhancement to clinical practice. Indeed, based on the opportunities described earlier in this chapter, were the DRAWN model to be employed, its use should result in more successful outcomes, and ANNs may well become ubiquitous to clinical practice, not unlike the technological device we call a stethoscope. Like ANN technology, the stethoscope, invented by René Laennec in 1819, was first resisted by many practitioners, but it was ultimately adopted as an essential tool of practice (Reiser, 1978, p. 29; Simmons, 2002, p. 65).

Limitations

Several limitations of this study present opportunities for future research. First, the number of studies that provided sufficient detail to address the required RQ responses was far smaller than originally anticipated. Reports of clinical research studies involving ANNs or other technology tools should provide a more robust description and specific detail as to both the software and the technical design of the tool so that future systematic reviews can extract this detail for analyses. If the management-scholar-practitioner community is expected to convert research findings (such as those gained in a systematic review like the present study) into practice, then the inclusion of technology details will be essential. Moreover, with the application of new technologies to practice at the forefront of that review effort, the motivation for the academic press to support inclusion of such critical detail has clearly been established by this work.

This review also restricted analysis to clinical studies that used ANN tools as primary to the research. The ANN is one of several such technology-based tools that can be applied to clinical care, with the literature including such applications as various Decision Tree systems,

Support Vector Machines, Random Forest Analysis, and *k*-Nearest Neighbor tools (all of which are, like ANNs, machine-learning systems). There are opportunities for the exploration of performance for many of those technological tools, whether in comparison to traditional approaches, or directly, as was done in this review.

Finally, the lack of scholarly literature on the means to manage the financial implications of new innovative and disruptive technologies in healthcare is challenging. It is difficult to provide the reader with some direct measure of economic viability when employing such software technologies, like AI applications, in healthcare practice. As was noted in Chapter 1, a return-on-investment analysis is simply not practical within the healthcare industry when the service the technology provides is not a direct diagnostic/prognostic tool (such as an MRI) that is subject to a service or procedure charge. Within healthcare, the use of a technology like an ANN is not typically identified as a revenue-generating (i.e., billable) activity that is processed as units of service applied against revenue generated; rather, it is implemented as a tool for error avoidance and risk mitigation, both of which are part of overhead expense and as such are more difficult to quantify in terms of return on investment. Thus, from the perspective of this review, it was problematic to determine a method to quantify the value of the ANN technology and thus to justify, on a purely economic basis, its implementation for diagnostic and prognostic support of clinical practice. Perhaps, given a different cost/revenue structure than is currently used in the U.S. healthcare system, that analysis might be made more clear and direct. In the meantime, this study affirms that ANN technology is valuable to practice when clinical outcomes can be measurably improved, even if that value is not directly expressed in economic terms. Suffice it to say, in clinical practice it is difficult to place a monetary value on “the patient lived.”

The Healthcare Manager’s Point of View

There is also further managerial consideration requiring analysis during implementation of the DRAWN model. As noted in Chapter 3, providing the ANN outcome to the clinician has some workflow timing considerations:

- 1) ANN outcomes presented too early can become satisficers.
- 2) ANN outcomes presented too late can cause a confirmation bias response.
- 3) The ANN outcome, therefore, must be treated like a Gladwell (2002) “tipping point” to the clinical decision, and its precise timing and delivery could itself influence the performance.

Thus, certainly more research on that tipping point is needed, as discussed in Chapter 5, in order for the ANN outcome presentation to be adjusted to meet the clinical workflow and the practice circumstances. This is a topic that the healthcare manager is likely best suited to evaluate on an individual practice basis, at least until further study is done and those results are made available.

Future Research Considerations

Additional research considerations, expanding from Chapter 5, include greater understanding of the clinical applications that lend themselves to ANN use and implementation.

This review identified breast cancer diagnostic assessment as one possibility, but those results were not strong due to the limited number of relevant studies available within this review.

However, one of the ANN systematic reviews mentioned as a model for this study was done on a similarly focused topic (Lisboa & Taktak, 2006). In their conclusions they found evidence in the literature to support ANN’s role in cancer diagnosis, with the caveat, however, that more work was needed to validate the reviewed studies’ findings using more traditional statistical approaches. This was a major design motivator for the DRAWN model, to maintain the clinician as the key decision maker when using automated systems like ANNs. We recognized that,

unlike physicians, ANNs are typically unable to contextualize findings, and that limitation would suggest the research would benefit not only from comparisons to traditional statistics (as this review had done) but in expanding larger scale application through piloting programs as a means to further evidence ANNs' validity and their flaws, as noted earlier. Thus, we look at the future of research into ANN use in healthcare as, minimally, a two-pronged approach – requiring a more stringent analysis of findings within the research body of evidence, *and* piloting usage under practical conditions to evaluate real-world outcomes.

And future considerations for research should also include non-ANN machine-learning tools, again as brought forth in Chapter 5. Just as there should be a two-pronged research approach for ANN tools, there are other Bayesian machine-learning systems that have been applied in the research literature that beg for similar analyses as presented here for ANNs. Indeed, within this study decision tool comparisons were noted but we did not evaluate the ANN alternatives themselves, only the relation of their performance against that of the ANN tool, and noting only whether the ANN fared better, worse, or equal to that competitor. Gleaning those comparative studies from the present data might give some insight into what other tools might be candidates for similar evaluation.

DRAWN Implications beyond Healthcare

Finally, although we have examined ANN application strictly within the healthcare industry, ANNs have been employed in a wide array of industries. It is instructive to ANN use and adoption to examine problems encountered in those alternative industries and to determine the applicability of those problems to healthcare use. A recent online article was published that seemed to illustrate the problems of implementing AI in general. Vincent (2016) claimed that there were three inhibitors to AI implementation – the need for extensive and available data for

learning, the ability for AI systems to multitask (with contextualization), and to uncover the learning process to identify limits. It is this last point that seems to have touched upon a shared issue in healthcare – the “black box” problem mentioned earlier. The article highlighted a case where an imaging system built to replicate an eye was given many images of rooms, with bedrooms specifically identified, to train it on what a bedroom looked like. It was then asked to identify what was covering the windows in the next (test) image. The computer responded correctly (curtains) but under the covers of the software’s process they found that once the computer found a bed in the room then “curtains” was the only, and obvious, answer. It turned out that all of the bedroom images it was shown in training had window curtains, so finding the bed in the test case was all it needed to know. While that is logical given what the computer was shown, it is certainly not the exclusive possibility in the real world – yet, to the computer, the “real” world only extends as far as its training data takes it.

In a healthcare setting this could relate to an ANN’s diagnostic suggestion of cancer if, given that all the training cases shown to the system for female patients in their 40s happened to have a confirmed cancer diagnosis, that particular case was for a female patient in her 40s. While that example may seem extreme, it, or a more complex version of it, is certainly within the realm of possibility. That also brings in to play Vincent’s (2016) other two problems – the availability of data (where were the healthy females who were in their 40s?) and the contextualization (the flawed logic of “Roses are red, the flower I have is red, hence it is a rose.”). These all strongly point to having a complex, multitasking, contextualizing analytical machine as an arbiter to the AI system’s decision – that is, it must pass the evaluative ability of a skilled human before action is taken. This seems to confirm the DRAWN model design, that the final decision is that of the human clinician and not simply accepting the ANN’s outcome.

The Future of AI Use in Healthcare

While it appears that the research literature is still relatively young regarding actual analysis of the practical (non-research) use of AI tools in healthcare, and of ANNs specifically, there is a wide range of evidence supporting *interest* in those tools for future healthcare applications. A simple Google search of the expression “future of artificial intelligence in healthcare” has revealed over 1.5 million hits (last done on 9/15/2017), and the trade press content seems quite extensive. Mesko, a physician blogger (*The Medical Futurist*), interestingly suggested that AI is the stethoscope of the 21st century (Mesko, 2017), noting similar arguments to those posed within this paper.¹¹ However, one business blogger (Weeks, n.d.) attributes the lack of AI use across all industries as related to the commensurate lack of meaningful access to all forms of data analytics – he posits that we are not trained to engage these tools in practice. A CNBC report (Choudhury, 2017) noted that AI can be a game-changer for practice, but getting clinicians onboard is certainly a tricky proposition, again, noting many of the same types of challenges to adoption examined here in this paper.

It is suggested, therefore, that the interest (e.g., the “buzz”) is certainly great for ANNs and other AI technologies to be incorporated into practice, and that the desire for such technologies appears strong as a means to improve human practice outcomes. Like the initial case for the stethoscope, resistance to adoption seems fairly strong, if the non-academic literature is to be believed, and yet there is in those writings a conviction that its eventual adoption seems inevitable, at least in some form. It is within this latter sphere that this paper can give a measure of guidance – piloting ANN tools using a DRAWN model of implementation might be one of the

¹¹ This article was discovered only *after* my argument was developed and written, although his was published online on July 10, 2017 (parallel, yet independent, models).

means of demonstrating effectiveness and value to the industry. What is certainly clear, however, is that *some* inherent changes to practice involving AI seem almost assured – less a case of *if* but one of *how*. The research presented here, it is hoped, will provide the mechanism, as well as the spark, needed to move that forward, and to give the practitioner an evidenced-based review to support practice change through AI tool integration.

Conclusions and the Future of Healthcare Practice

The examination of ANN performance through the two RQs has provided the “what works” answer, both in the overall sense (RQ #1) as well as, to some lesser degree, where it works best (RQ #2). This paper includes a rationale for changes to practice using the DRAWN model proposed, and identifies the capabilities that such an implementation could bring to bear upon clinical practice. Given the performance of ANNs and humans independently, the model suggests that appropriately combining the two within a workflow process will be productive from a quality perspective, and, as a consequence of that, very likely from a financial perspective as well. Given the model as proposed, this still leaves the human clinician at the center of the decision; this acknowledges that the scope and depth of a human brain is still well beyond the reach of ANN technology, at least for the foreseeable future.

Finally, in the context of the AOM topics selected for this study (Chapter 1), the following were addressed in this review as noted:

- 1) Healthcare Management – Use of ANNs in diagnostic and prognostic assessment has been demonstrated to improve the performance of healthcare workers and organizations through better quality of care and improved patient outcomes, with associated implications for better management of healthcare organizations.

- 2) Technology and Innovation Management – As a disruptive innovation to clinical practice, the ANN is a technology whose development, implementation, and use in technically-oriented activities (diagnostic and prognostic assessment) can be integrated into the organization and its practice.

So what could the future of healthcare look like with widespread adoption of ANN technologies through the DRAWN model implementation? Beyond simply providing individual ANNs to practitioners, it is possible that libraries of these tools can be managed, offered (through subscription), and updated by follow-on research. As noted earlier, Web applications could be either made available as stand-alone products or integrated with EMR systems whose vendors engage the ANN library through a standardized interface. And as the use of EMRs expands, so would the availability of this ANN clinical consultant, assisting providers in remote locations around the globe as well as those in poor rural and urban venues that lack the kinds of resources for clinical consultancy that our government's own healthcare payer, CMS, is demanding.

Future research as suggested here can bear great fruit for future generations of clinical managers in enhancing and expanding advanced clinical practice around the nation, and the world, making both better places for it.

References

- Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Thousand Oaks, CA: SAGE Publications, Inc.
- Academy of Management. (2007). *Academy of Management division & interest group domain statements: Division and interest group domains*. Retrieved from <http://aom.org/Divisions-and-Interest-Groups/Academy-of-Management-Division---Interest-Group-Domain-Statements.aspx>
- Adya, M., & Collopy, F. (1998). How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, 17(5-6), 481-495. doi:10.1002/(sici)1099-131x(1998090)17:5/6<481::aid-for709>3.0.co;2-q
- Alsing, S. G., Bauer, K. W., & Oxley, M. E. (2002). Convergence for receiver operating characteristic curves and the performance of neural networks. *International Journal of Smart Engineering System Design*, 4(2), 133-145. doi:10.1080/10255810290008054
- Amaral, A., & Krishna, V. (2017, August 1). HIT Think: Why robotics and AI still face an uphill battle in healthcare [Opinion]. Retrieved from <https://www.healthdatamanagement.com/opinion/>
- Amato, F., López, A., Peña-Méndez, E. M., Vañhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis [Editorial]. *Journal of Applied Biomedicine*, 11(2), 47-58. doi:10.2478/v10136-012-0031-x
- American Reinvestment and Recovery Act of 2009, Public Law 111-5, 123 Statute 247-248
- Andersson, S., Heijl, A., Bizios, D., & Bengtsson, B. (2013). Comparison of clinicians and an artificial neural network regarding accuracy and certainty in performance of visual field

assessment for the diagnosis of glaucoma. *Acta Ophthalmologica*, 91(5), 413-417.

doi:10.1111/j.1755-3768.2012.02435.x

Atawande, A. (2009). *The Checklist Manifesto: How to get things right*. New York, NY: Henry Holt & Co.

Balogh, E., Miller, B. T., & Ball, J. (Eds.). (2015). *Improving Diagnostics in Health Care*.

Washington, DC: The National Academies Press. doi:10.17226/21794

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4), 387-415.

doi:10.1016/0022-2496(75)90001-2

Barends, E., ten Have, S., & Huisman, F. (2012). Learning from other evidence-based practices:

The case of medicine. In D. Rousseau (Ed.), *The Oxford Handbook of evidence-based management* (pp. 25-42). Oxford, UK: Oxford University Press.

Barnett-Page, E., & Thomas, J. (2009). Methods for the synthesis of qualitative research: a

critical review. *BMC Medical Research Methodology*, 9(59). doi:10.1186/1471-2288-9-59

Bartosch-Härlid, A., Andersson, B., Aho, U., Nilsson, J., & Andersson, R. (2008). Artificial

neural networks in pancreatic disease. *The British Journal of Surgery*, 95(7), 817-826.

doi:10.1002/bjs.6239

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing,

design, and application. *Journal of Microbiological Methods*, 43(1), 3-31.

Bate, L., Hutchinson, A., Underhill, J., & Maskrey, N. (2012). How clinical decisions are

made. *British Journal of Clinical Pharmacology*, 74(4), 614-620. doi:10.1111/j.1365-

2125.2012.04366.x

Baxt, W. G., & Skora, J. (1996). Prospective validation of artificial neural network trained to identify acute myocardial infarction. *Lancet*, 347(8993), 12-15. doi:10.1016/S0140-6736(96)91555-X

Bazerman, M. H., & Moore, D. A. (2013). *Judgment in Managerial Decision Making* (8th Ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Belliveau, T., Axt, J., & Seetharama, M. (2015, November). *Informatics support for diagnostic classification and prognostic estimation, demonstrated with neurotrauma data*. Lecture/presentation given at the 29th Annual Convention of the Connecticut Psychology Association, Haddam, CT.

Belliveau, T., Seetharama, M., & Axt, J. (2015a, April). *Optimal prediction of spinal cord injury outcomes: Progress and challenges developing clinical decision support tools based on artificial neural network analyses of the National SCI Model Systems data*. Presentation given at to staff at the New England Regional Spinal Cord Injury Center, School of Public Health, Boston University, Boston, MA.

Belliveau, T., Seetharama, M., & Axt, J. (2015b, June). *Using an Artificial Neural Network for prognostic decision support in Spinal Cord Injury*. Lecture/presentation given at the annual Research Day Symposium at Hospital for Special Care, New Britain, CT.

Belliveau, T., Jette, A. M., Seetharama, S., Axt, J., Rosenblum, D., Larose, D.,...Larose, C. (2016). Developing artificial neural network models to predict functioning one year after traumatic spinal cord injury. *Archives of Physical Medicine and Rehabilitation*, 97(10), 1663-1668. doi:10.1016/j.apmr.2016.04.014

- Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5), S2-S23.
doi:10.1016/j.amjmed.2008.01.001
- Bethel, E. C., & Bernard, R. M. (2010). Developments and trends in synthesizing diverse forms of evidence: Beyond comparisons between distance education and classroom instruction. *Distance Education*, *31*(3), 231-256.
- Bigus, J. P. (1996). *Data mining with neural networks*. New York, NY: McGraw-Hill
- Blumenthal-Barby, J. S., & Burroughs, H. (2012). Seeking better health care outcomes: The ethics of using the “nudge”. *The American Journal of Bioethics*, *12*(2), 1-10.
doi:10.1080/15265161.2011.634481
- Briner, R. B. & Denyer, D. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. In D. Rousseau (Ed.), *The Oxford Handbook of evidence-based management* (pp. 112-129). Oxford, UK: Oxford University Press.
- Briner, R. B., Denyer, D., & Rousseau, D. M. (2009). Evidence-based management: Concept cleanup time? *Academy of Management Perspectives*, *23*(4), 19-32.
doi:10.5465/AMP.2009.45590138
- Burke, T. (2010). The health information technology provisions in the American Recovery and Reinvestment Act of 2009: Implications for public health policy and practice. *Public Health Reports*, *125*(1), 141-145.
- Burns, R. B., & Burns, R. A. (2012). *Business and research methods and statistics using SPSS*. London, UK: SAGE Publications Ltd.
- Burnes, B. (2004). Kurt Lewin and the planned approach to change: A re-appraisal. *Journal of Management Studies*, *41*(6), 977-1002. doi:10.1111/j.1467-6486.2004.00463.x

- Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine [Opinion]. *Journal of the American Medical Association*, 318(6), 517-518.
- Campitelli, G., & Gobet, F. (2010). Herbert Simon's decision-making approach: Investigation of cognitive processes in experts. *Review of General Psychology*, 14(4), 354-364.
doi:10.1037/a0021256
- Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference: Machine Learning*, 161.
doi:10.1145/1143844.1143865
- Caudill, M., & Butler C. (1990). *Naturally intelligent systems*. Cambridge, MA: Massachusetts Institute of Technology.
- Centers for Medicare and Medicaid Services. (2016). *CMS Quality Strategy Update 2016*.
Retrieved from <http://www.cms.gov>
- Centers for Medicare and Medicaid Services. (2017). *National health expenditure data: Historical*. Retrieved from <https://www.cms.gov/>
- Christensen, C. M. (1997). *The Innovator's Dilemma: When new technologies cause great firms to fail*. Boston, MA: Harvard Business Review Press.
- CMS.gov. (2017). CMS.gov: Centers for Medicare and Medicaid Services (Home Page).
Retrieved from <https://www.cms.gov/index.html>
- Collopy, F., Adya, M., & Armstrong, J. S. (1994). Principles for examining predictive validity: The case of information systems spending forecasts. *Information Systems Research*, 5(2), 170-179. doi:10.1287/isre.5.2.170
- Coye, M. J., & Kell, J. (2006). How hospitals confront new technology. *Health Affairs*, 25(1), 163-173. doi: 10.1377/hlthaff.25.1.163

Croskerry, P. (2002). Achieving quality in clinical decision making: Cognitive strategies and detection of bias. *Academic Emergency Medicine*, 9(11), 1184-1204.

doi:10.1197/aemj.9.11.1184

Croskerry, P. (2014). Bias: A normal operating characteristic of the diagnosing brain. *Diagnosis*, 1(1), 23-27. doi:10.1515/dx-2013-0028

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837-845. doi:10.2307/2531595

Dent, E. B., & Powley, E. H. (2003). Employees actually embrace change: The chimera of resistance. *Journal of Applied Management and Entrepreneurship*, 8(1), 40-56. Retrieved from

<http://ezproxy.umuc.edu/login?url=http://search.proquest.com/docview/203904216?accountid=14580>

Denyer, D. & Tranfield, D. (2009). Producing a systematic review. In D.A. Buchanan & A. Bryman (Eds.), *The Sage Handbook of Organizational Research Methods* (pp. 671-689). Los Angeles: Sage Publications.

Donaldson, L. (2012). Evidence-based management (EBMgt) using organizational facts. In D. Rousseau (Ed.), *The Oxford Handbook of evidence-based management* (pp. 249-261). Oxford, UK: Oxford University Press.

Dybowski, R., & Gant, V. (2001). *Clinical applications of artificial neural networks*. Cambridge, MA: Cambridge University Press.

- Faisal, T., Taib, M., & Ibrahim, F. (2012). Neural network diagnostic system for dengue patients risk classification. *Journal of Medical Systems*, 36(2), 661-676. doi:10.1007/s10916-010-9532-x
- Ferreira, P. R., Ferreira, R. F., Rajgor, D., Shah, J., Menezes, A., & Pietrobon, R. (2010). Clinical reasoning in the real world is mediated by bounded rationality: Implications for diagnostic clinical practice guidelines. *PLoS One*, 5(4), 1-8. doi:10.1371/journal.pone.0010265
- Flyvbjerg, B. (2004). Five misunderstandings about case research. In C. Seale, G. Gobo, J. F. Gubrium, & D. Silverman (Eds.), *Qualitative Research Practice*, (1st ed., pp. 420-434). Thousand Oaks, CA: Sage.
- Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. Thousand Oaks, CA: Sage Publications Ltd.
- Gant, V., Rodway, S., & Wyatt, J. (2001). Artificial neural networks: Practical considerations for clinical application. In R. Dybowski & V. Gant, *Clinical applications of artificial neural networks* (pp. 329-356). Cambridge, MA: Cambridge University Press.
- Ghavami, P. K. (2012). *An investigation of applications of artificial neural networks in medical prognostics* [Doctoral dissertation]. Retrieved from ProQuest Dissertations and Theses. (Order No. 3542349, University of Washington).
- Gladwell, M. (2002). *The Tipping Point: How little things can make a big difference*. New York, NY: Little, Brown and Company.
- Goode, L., Clancy, C., Kimball, H., Meyer, G., & Eisenberg, J. (2002). When is "good enough"? The role and responsibility of physicians to improve patient safety. *Academic Medicine*, 77(10), 947-952. doi:10.1097/00001888-200210000-00004

- Gough, D. (2007). Weight of Evidence: a framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, 22(2), 213-228.
doi:10.1080/02671520701296189
- Gough, D., Oliver, S., and Thomas, J. (Eds.). (2012). *An introduction to systematic reviews*. London, UK: SAGE Publications Ltd.
- Graber, M. L., & Carlson, B. (2011). Diagnostic error: The hidden epidemic. *Physician Executive*, 37(6), 12-19. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22195411>
- Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine*, 165(13), 1493-1499.
doi:10.1001/archinte.165.13.1493
- Grajczyk, K. (2008). *Designing and exploring intelligent decision support systems: A description of five technologies and an implementation case study for an Artificial Neural Network* [Kindle Fire version]. Retrieved from Amazon.com.
- Harden, A., & Gough, D. (2012). Quality and relevance appraisal. In D. Gough, S. Oliver, and J. Thomas (Eds.), *An introduction to systematic reviews* (pp. 153-178). London, UK: SAGE Publications Ltd.
- Harden, A., & Thomas, J. (2005). Methodological Issues in Combining Diverse Study Types in Systematic Reviews. *International Journal of Social Research Methodology*, 8(3), 257-271. doi:10.1080/13645570500155078
- Hawkins, R. C. (2007). Laboratory turnaround time. *Clinical Biochemist Reviews*, 28(4), 179-194. Retrieved from <http://www.aacb.asn.au/clinical-biochemist-reviews>

- Hess, E. P., Thiruganasambandamoorthy, V., Wells, G. A., Erwin, P., Jaffe, A. S., Hollander, J. E., ...Stiell, I. G. (2008). Diagnostic accuracy of clinical prediction rules to exclude acute coronary syndrome in the emergency department setting: A systematic review. *Canadian Journal of Emergency Medicine, 10*(4). 373-382.
- Hogg, R. V., Tanis, E. A., & Zimmerman, D. (2015). *Probability and Statistical Inference (9th Ed.)*. Upper Saddle River, NJ: Pearson Education, Inc.
- Holst, H., Ohlsson, M., Peterson, C., & Edenbrandt, L. (1999). A confident decision support system for interpreting electrocardiograms. *Clinical Physiology, 19*(5), 410-418.
doi:10.1046/j.1365-2281.1999.00195.x
- Institute of Medicine. (1999). *To Err is Human: Building a safer health system*. Washington, DC: National Academy Press.
- Institute of Medicine. (2012). *Best Care at Lower Cost: The path to continuously learning health care in America*. Washington, DC: National Academy Press.
- Jacob, M., Lewsey, J., Sharpin, C., Gimson, A., Rela, M., & van der Meulen, J. (2005). Systematic review and validation of prognostic models in liver transplantation. *Liver Transplantation, 11*(7), 814-825. doi:10.1002/lt.20456
- Jao, C. S., & Hier, D. B. (2010). Clinical decision support systems: An effective pathway to reduce medical errors and improve patient safety. In C. S. Jao, D. B. Hier (eds.), *Decision Support Systems* (pp. 121-138). Retrieved from:
<http://www.intechopen.com/books/decision-support-systems/clinical-decision-support-systems-an-effective-pathway-to-reduce-medical-errors-and-improve-patient-safety>
doi:10.5772/39469

- Jaspers, M. W. M., Smeulers, M., Vermeulen, H., & Peute, L. W. (2011). Effects of clinical decision-support systems on practitioner performance and patient outcomes: A synthesis of high-quality systematic review findings. *Journal of the American Medical Informatics Association*, 18(3), 327-334. doi:10.1136/amiajnl-2011-000094
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
- Kohonen, T. (2001). *Self-Organizing Maps*, (3rd E.). New York, NY: Springer-Verlag.
- Kukull, W., & Ganguli, M. (2012). Generalizability - The trees, the forest, and the low-hanging fruit. *Neurology*, 78(23), 1886-1891. doi:10.1212/WNL.0b013e318258f812
- Kumar, S. S., & Kumar, K. A. (2013). Neural networks in medicine and healthcare. *International Journal of Innovative Research and Development*, 2(8), 241-244. Retrieved from <http://www.ijird.com>
- Lan, J. (2005). *Asymmetric misclassification costs and imbalanced group sizes in neural networks for classification* [Doctoral dissertation]. Retrieved from ProQuest Dissertations and Theses. (Order No. 3184002, Kent State University).
- Lancashire, L. J., Lemetre, C., & Ball, G. R. (2009). An introduction to artificial neural networks in bioinformatics application to complex microarray and mass spectrometry datasets in cancer studies. *Briefings in Bioinformatics*, 10(3), 315-329. doi:10.1093/bib/bbp012
- Lee, H., Hwang, S., Han, S., Park, S., Kim, S., Cho, J., & ... Choe, G. (2010). Image-based clinical decision support for transrectal ultrasound in the diagnosis of prostate cancer: comparison of multiple logistic regression, artificial neural network, and support vector machine. *European Radiology*, 20(6), 1476-1484. doi:10.1007/s00330-009-1686-x
- Lee, K. (2017, June 24). The real threat of artificial intelligence [Sunday Review - Opinion]. *The New York Times*, Retrieved from <https://nyti.ms/2t3gJIU>

- Lewin, K. (1947). Frontiers in group dynamics: Concept, method and reality in social science, social equilibria and social change. *Human Relations*, 1(1), 5-41. doi: 10.1177/001872674700100103
- Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A.,...Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Annals of Internal Medicine*, 151(4), W-65-W-94. doi:10.7326/0003-4819-151-4-200908180-00136
- Liebowitz, J. (2001). Knowledge management and its link to artificial intelligence. *Expert Systems with Applications*, 20(1), 1-6. doi: 10.1016/s0957-4174(00)00044-0
- Lisboa, P. J. G. (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks*, 15(1), 11-39. doi:10.1016/S0893-6080(01)00111-3
- Lisboa, P. J., & Taktak, A. G. (2006). The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks*, 19(4), 408-415. doi:10.1016/j.neunet.2005.10.007
- Litwin, A. (2011). Technological change at work: The impact of employee involvement on the effectiveness of health information technology. *Industrial & Labor Relations Review*, 64(5), 863-888.
- M*Modal. (2017). CDI Engage. Retrieved from <https://mmodal.com/>
- Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, 90, 46-60. doi: 10.1016/j.futures.2017.03.006

- Mamede, S., van Gog, T., van den Berge, K., Rikers, R., van Saase, J., van Guldener, C., & Schmidt, H. (2010). Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *Journal of the American Medical Association*, 304(11), 1198-1203. doi:10.1001/jama.2010.1276
- Marshall, H. S., & Milikowski, C. (2017). Comparison of clinical diagnoses and autopsy findings: Six-year retrospective study [Manuscript – early online release]. *Archives of Pathology & Laboratory Medicine In-Press*. doi:10.5858/arpa.2016-0488-OA
- Marion, J. (2016, October 13). Realities of man versus machine diagnosis [Web log posting]. Retrieved from <https://www.healthcare-informatics.com/blogs/>
- Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of The Royal Meteorological Society*, 128(584), 2145-2166. doi:10.1256/003590002320603584
- Mays, N., Pope, C., & Popay, J. (2005). Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of Health Services Research & Policy*, 10(S1), 106-120. doi:10.1258/1355819054308576
- McLaren, C. E., Chen, W., Nie, K., & Su, M. (2009). Prediction of malignant breast lesions from MRI features: a comparison of artificial neural network and logistic regression techniques. *Academic Radiology*, 16(7), 842–851. doi:10.1016/j.acra.2009.01.029
- McMillan, N. (2017, August 24). HIT Think: How machine learning can speed quality measure development [Opinion]. Retrieved from <https://www.healthdatamanagement.com/opinion/>

Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J. M., Maino, K., . . . Hamann, J.

(2011). Confirmation bias: Why psychiatrists stick to wrong preliminary

diagnoses. *Psychological Medicine*, 41(12), 2651-2659.

doi:<http://dx.doi.org/10.1017/S0033291711000808>

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook*. Thousand Oaks, CA: SAGE Publications, Inc.

Müller, B., Reinhardt, J., & Strickland, M. T. (1995). *Neural networks: An introduction* (2nd Ed.). Berlin, DE: Springer-Verlag.

Nehme, Z., Boyle, M., & Brown, T. (2013). Diagnostic accuracy of prehospital clinical prediction models to identify short-term outcomes in patients with acute coronary syndromes: A systematic review. *Journal of Emergency Medicine*, 44(5), 946-954.

doi:10.1016/j.jemermed.2012.07.078

NLM/NIH PubMed Central. (2017). *PMC: Authors Manuscript in PMC*. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/>

Oliver, S., Dickson, K., and Newman, M. (2012). Getting started with a review. In D. Gough, S. Oliver, and J. Thomas (Eds.), *An introduction to systematic reviews* (pp. 66-82). London, UK: SAGE Publications Ltd.

Oliver, S. & Sutcliffe, K. (2012). Describing and analysing studies. In D. Gough, S. Oliver, and J. Thomas (Eds.), *An introduction to systematic reviews* (pp. 135-152). London, UK: SAGE Publications Ltd.

Olsson, S., Ohlsson, M., Ohlin, H., Dzaferagic, S., Nilsson, M., Sandkull, P., & Edenbrandt, L. (2006). Decision support for the initial triage of patients with acute coronary

- syndromes. *Clinical Physiology and Functional Imaging*, 26(3), 151-156. doi: 10.1111/j.1475-097x.2006.00669.x
- Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, 56(1), 45-50. doi:10.4103/0301-4738.37595
- Patel, J. L., & Goyal, R. K. (2007). Applications of artificial neural networks in medical science. *Current Clinical Pharmacology*, 2(3), 217-226. doi:10.2174/157488407781668811
- Pearson, C. M., & Clair, J. A. (1998). Reframing crisis management. *Academy of Management Review*, 23(1), 59-76. doi:10.5465/AMR.1998.192960
- Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden, MA: Blackwell Publishing.
- Pham, J., Aswani, M., Rosen, M., Lee, H., Huddle, M., Weeks, K., & Pronovost, P. (2012). Reducing Medical Errors and Adverse Events. *Annual Review of Medicine*, 63(1), 447-463. doi:10.1146/annurev-med-061410-121352
- Pilon, M. (2015, November 30). *Doc migration: Many CT-minted physicians leave state*. Retrieved from <http://www.hartfordbusiness.com/article/20151130/PRINTEDITION/311259935/>
- Price Waterhouse Coopers. (2010). The price of excess: Identifying waste in health care spending. Retrieved from <http://www.pwc.com/us/en/healthcare/publications/the-price-of-excess.html>

Recovery.gov. (2014, October 11). *The American Recovery and Reinvestment Act – Advanced Recipient Data Search*. Retrieved from

<http://www.recovery.gov/arra/espsearch/Pages/advanced.aspx>

Reeves, B. C., Deeks, J. J., Higgins, J. P. T., & Wells, G. A. (2011). Including non-randomized studies (Chapter 13). In J. P. T. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Review of Interventions, Version 5.1.0*. Retrieved from

www.handbook.cochrane.org

Reio, T. G. (2009). Contributing to the emergent research method conversation. *Human Resource Development Quarterly*, 20(2), 143-146. doi:10.1002/hrdq.20012.

Reiser, S. J. (1978). *Medicine and the Reign of Technology*. Boston, MA: Cambridge University Press.

Rometty, G. (2017, February 20). *HIMSS17 Keynote Speaker: Ginni Rometty*. Retrieved from <https://www.youtube.com/watch?v=D58bqSJr6Mg>

Rousseau, D. M. (2006). 2005 Presidential Address: Is there such a thing as 'Evidence-Based Management'?. *The Academy of Management Review*, 31(2), 256-269.
doi:10.2307/20159200

Rousseau, D. (2012). Envisioning evidenced-based management. In D. Rousseau (Ed.), *The Oxford Handbook of evidence-based management* (pp. 3-24). Oxford, UK: Oxford University Press.

Rousseau, D. M., Manning, J., & Denyer, D. (2008). Chapter 11: Evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses. *Academy Of Management Annals*, 2(1), 475-515.
doi:10.1080/19416520802211651

- Rumsfeld, D. (2002, June 6). *Press conference: NATO HQ, Brussels*. Retrieved from <http://www.nato.int/docu/speech/2002/s020606g.htm>
- Russell, S., & Norvig, P. (2014). *Artificial Intelligence: A Modern Approach (3rd Ed.)*. Essex, UK: Pearson Education Limited.
- Simmons, J. G. (2002). *Doctors and Discoveries: Lives that created today's medicine*. Boston, MA: Houghton Mifflin Co.
- Simon, H. A. (1943). *A theory of administrative decision* [Doctoral dissertation]. Retrieved from ProQuest Dissertations and Theses database. (Order No. T-07632).
- Simon, H. A. (1979). Rational Decision Making in Business Organization. *American Economic Review*, 69(4), 493-513.
- Simon, H. A. (1992). What is an "explanation" of behavior? *Psychological Science*, 3(3), 150-161. doi:10.1111/j.1467-9280.1992.tb00017.x
- Simon, H. A. (1997). *Administrative Behavior, 4th Ed.* New York, NY: The Free Press – Simon and Schuster, Inc.
- Smith, M., Saunders, R., Stuckhardt, L., & McGinnis, J. M. (Eds.). (2012). *Best care at lower cost: The path to continuously learning health care in America*. Washington, DC: The National Academies Press.
- Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5), 645-665. doi:10.1037/10315-017
- Stergiou, C., & Siganos, D. (1997, May 12). *Neural Networks*. Retrieved from http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

Steward, R. & Oliver, S. (2012). Making a difference with systematic reviews. In D. Gough, S. Oliver, and J. Thomas (Eds.), *An introduction to systematic reviews* (pp. 227-244).

London, UK: SAGE Publications Ltd.

Thaler, R. H. (2000). Interdisciplinary studies of consciousness: From homo economicus to homo sapiens. *Journal of Economic Perspectives*, 14(1), 133–141. doi:10.17323/1995-459x.2007.4.32.35

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New York, NY: Penguin Group, Inc.

Thammasitboon, S., & Cutrer, W. B. (2013). Diagnostic decision-making and strategies to improve diagnosis. *Current Problems in Pediatric and Adolescent Health Care*, 43(9), 232-241. doi:10.1016/j.cppeds.2013.07.003

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and bias. *Science*, 185(4175), 1124-1131. doi:10.1126/science.185.4157.1124

Uğuz, H. (2012). A Biomedical System Based on Artificial Neural Network and Principal Component Analysis for Diagnosis of the Heart Valve Diseases. *Journal of Medical Systems*, 36(1), 61-72. doi:10.1007/s10916-010-9446-7

UMUC Library. (2016). *About UMUC Library OneSearch* [Website]. Retrieved from <http://www.umuc.edu/library/libhow/onesearch.cfm#databases>

Van de Ven, A. H., & Poole, M. S. (1995). Explaining development and change in organizations. *Academy of Management Review*, 20(3), 510-540. Retrieved from <http://www.jstor.org/stable/258786>

- VanVoorhis, C. R., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, 3(2), 43-50. doi:10.20982/tqmp.03.2.p043
- Vincent, J. (2016, October 10). These are three of the biggest problems facing today's AI. *The Verge*. Retrieved from <https://www.theverge.com/2016/10/10/13224930/ai-deep-learning-limitations-drawbacks>
- Weick, K. E. (1993). The Collapse of Sensemaking in Organizations: The Mann Gulch Disaster. *Administrative Science Quarterly*, 38(4), 628-652. doi:10.2307/2393339
- Weick, K. (2005). The experience of theorizing: Sensemaking as topic and resource. In K. G. Smith & M. A. Hitt (Eds.), *Great minds in management: The process of theory development* (pp. 394-413). Oxford, UK: Oxford University Press.
- Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. *Organization Science*, 16(4), 409-421. doi:10.2307/25145979
- West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and Biases as Measures of Critical Thinking: Associations with Cognitive Ability and Thinking Dispositions. *Journal of Educational Psychology*, 100(4), 930-941. doi:10.1037/a0012842
- Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1), 35-62. doi:10.1016/S0169-2070(97)00044-7

Appendix A

An Overview of ANNs

Introduction

To begin, it must be acknowledged that this overview is intended to give the reader a general understanding of artificial neural network technology and its application, and not a detailed review of the theoretical and mathematical underpinnings of that technology. Our approach would be akin to a driver's understanding of an automobile – being able to effectively employ and maneuver it while not having a fully detailed understanding of the components of the power train or how the internal combustion engine at its core actually functions. It is understood that this may not appeal to the technologists in the reading audience, much as Garson (1998) had noted in his discussion of obstacles to ANN acceptance (quoting Professor Andrew Hunter, Dean of Science, Technology, and Engineering at University of Lincoln, UK):

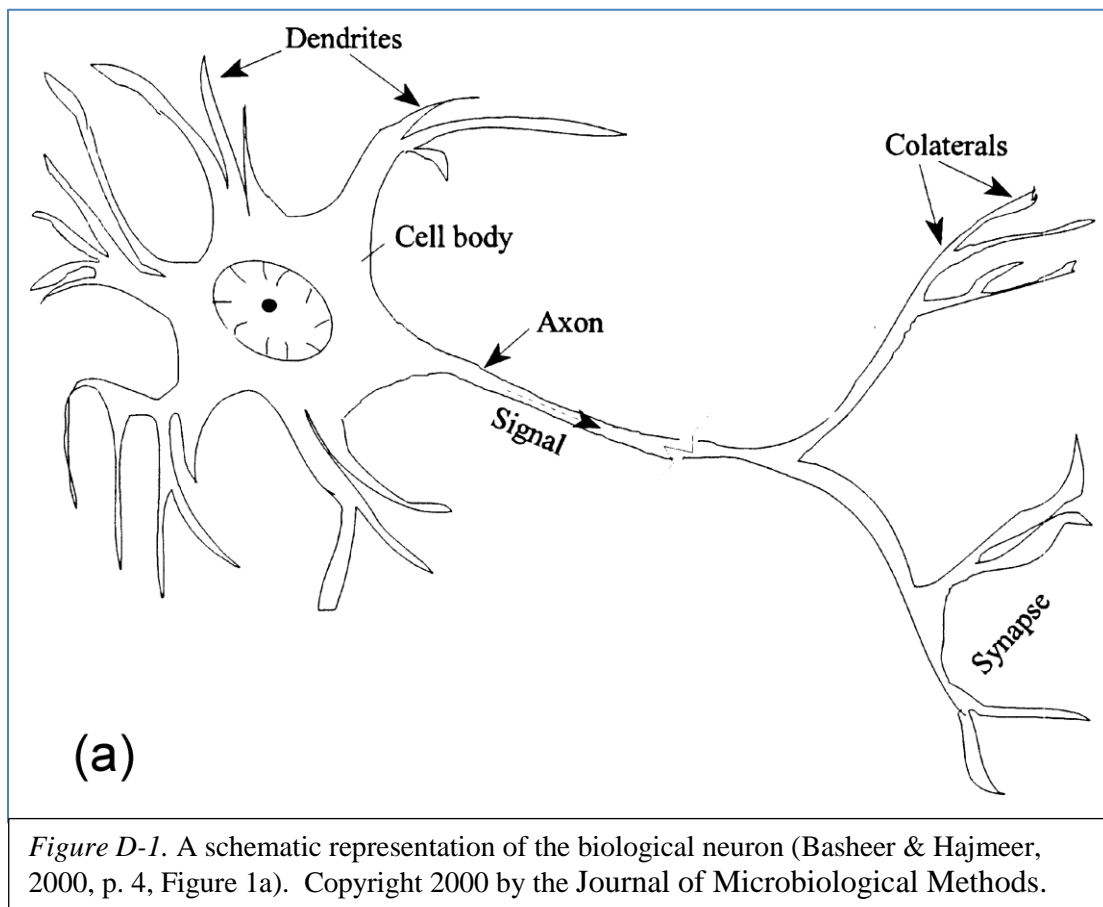
It is accepted that it may never be possible to completely understand how such a machine actually operates....From a viewpoint of a classically-trained scientist, building things without understanding them is almost tantamount to heresy. (Garson, 1998, p. 16)

Yet, given the limits posed by the complexity of the technology, we address this issue in as transparent a manner as would be possible without subjecting the reader to a full academic course on the technology, given that such transparency is tenet of systematic review (Briner, Denyer, & Rousseau, 2009, p. 23; Gough, 2007, p. 224; Harden & Gough, 2012, p. 157; Stewart & Oliver, 2012, p. 240). However, for those who wish to pursue further information regarding ANN function at a higher level of detail, you are referred to Müller et al.'s 1995 text, *Neural Networks: An Introduction*, on the physics of neural networks (Müller, Reinhardt, & Strickland, 1995).

Given that expressed limitation, we next endeavor to review the ANN, first structurally, then theoretically (from an application perspective), and finally, from a productivity point of view (how can they be actually employed to the benefit of the social scientist or manager).

ANN Structure

The model of an artificial neural network is essentially the biological one – the neurons that constitute the central nervous systems present in all higher-order life forms on Earth. The biological neuron is schematically pictured below (Figure D-1), with the major components – the cell body (or Soma), the dendrites (and related collaterals), and the axon – all labeled for clarity:



What is crucial to observe is the direction of “information” flow, from the dendrites through the cell body, and on to the axon, as that is the pathway the transmission takes through the cell (and

is represented by the “Signal” arrow at center). In a typical network the ends of the axon filaments terminate at, or extremely close to, the dendrites of other neurons, at a connection point called a synapse (Caudill & Butler, 1990, p. 14; Müller, Reinhart, & Strickland, 1995, pp. 3-7), creating a continuous web-like structure (see Figure D-2).

The ‘decision-maker’ of the neuron is within the cell body, where the dendrite inputs are collected and, if the input is strong enough, the neuron will push its signal down the axon in order to incite

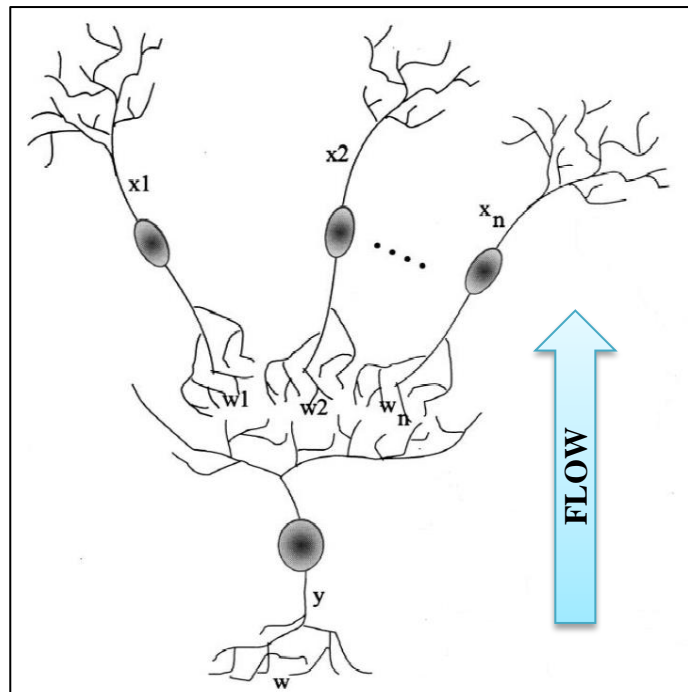
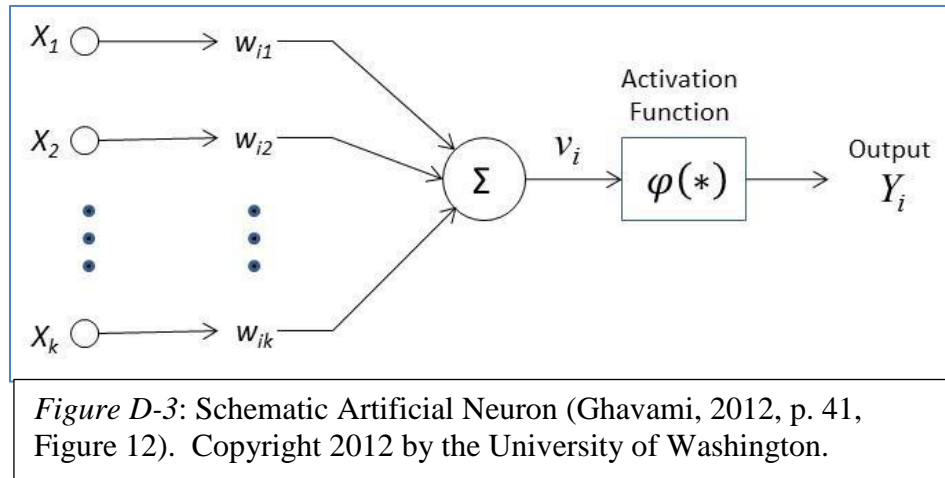


Figure D-2. A schematic representation of a biological neural network (Basheer & Hajmeer, 2000, p. 5, Figure 2). Copyright 2000 by the Journal of Microbiological Methods.

the next neuron (or neurons) in sequence (Caudill & Butler, 1990, pp. 14-15; Garson, 1998, pp. 23-24; Müller, Reinhart, & Strickland, 1995, pp. 4-6). One can image chains of these neural connections providing communication pathways along which information is carried, but that does not explain how the neural network learns. For this we turn to a more simplistic model – the artificial neuron.

We provide an analogous schematic representation (Figure D-3) to the neuron in Figure D-1, understanding that the components, while similar to a biological neuron, function somewhat differently. In Figure D-3 we indicate a series of input elements labeled X_k (equivalent to the “synapses”), pathways from those inputs to the body labeled w_{ik} (“dendrites”), the decision-making body itself (“soma”) – the combination of the Summarization (Σ) and Activation (φ)

functions – and the pathway V_i for the output signal to travel (“axon”) to the output Y_i . Thus, all the same components are there, but they function not by chemical interaction (like a synapse) or an electron-pump (the cell body) – rather they are simple digital models of these components which apply a decision rule to enforce what will happen when inputs are applied on the left, and whose action(s) determine what, if any, output is provided on the right.



Both the biological and artificial neurons have an Activation Function (what takes place in the body of a biological neuron) that determines whether or not the neuron should send an output (whether to fire at all in the case of the biological unit, or to pass along a summed value in artificial neurons). It can be noted that the inputs can be either “excitatory” (have a positive influence) or “inhibitory” (a negative influence) on whether the neuron is to pass the information along by firing, and it is the sum of those effects that influences the actual decision (hence the Σ function). In the biological version the chemical reactions at the synapse dictate whether the input is excitatory or inhibitory, but for the artificial neuron the process is different. It is the “w” factors, representing “weights” to be applied to the inputs, that provide that differentiating ability – the weights representing a value by which the input is adjusted (multiplied) in order to adjust each particular input’s strength (positive or negative, relating to excitatory or inhibitory). What

is also true of the artificial neuron is that adjusting those weights is the mechanism by which a neuron “learns” – the weight adjustment, by influencing the Activation Function, alters the neuron’s firing mechanism, hence its output.

This artificial neuron, as originally envisioned by McCulloch and Pitts in 1943 (Lancashire, Lemetre, & Ball, 2009, p. 316), came to be called a *perceptron*, as was noted in Chapter 1 of this paper, and represents the most simple core component of the ANN. The current perceptron model, based on a design by Rosenblatt in 1958 (Dybowski & Gant, 2001, pp. 2; Lan, 2005, p. 17; Stergiou & Siganos, 1997, Appendix A) generally has more than one input but only one output. Hence, only a simplistic level of decision-making can be enacted. However, when perceptrons are networked together in parallel (each input is applied to *all* of the perceptrons simultaneously) then decisions can be distributed across them. What is more, the ANN model prescribes not one “layer” of perceptrons, but at least three – an input layer, one or more so-called “hidden” layer(s), and an output layer. Thus, the network builds as noted in the schematic shown as Figure D-4:

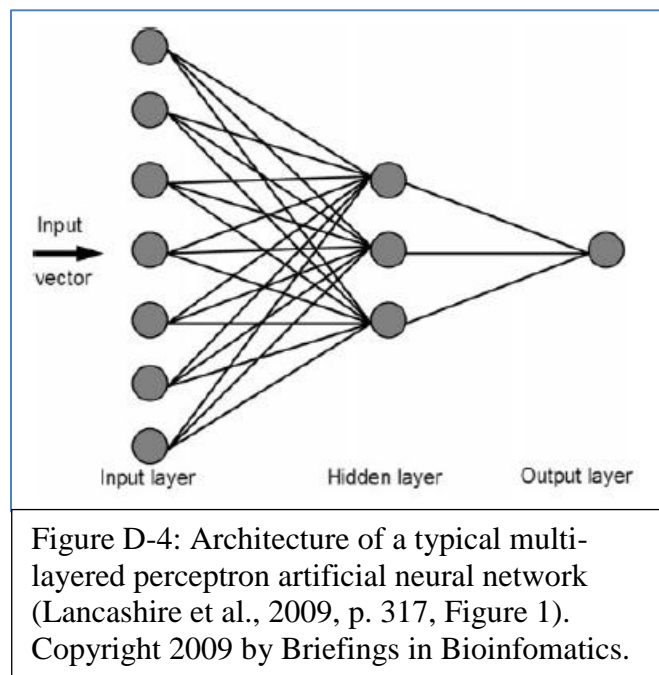


Figure D-4: Architecture of a typical multi-layered perceptron artificial neural network (Lancashire et al., 2009, p. 317, Figure 1). Copyright 2009 by Briefings in Bioinformatics.

The input layer interacts with the external environment to receive the input data as a vector of independent variables, each represented as a node. This information is passed through the first hidden layer, and multiplied (thus modified) by a set of associated weights. These products are summed and fed through a non-linear transfer function (e.g., sigmoid or hyperbolic tangent) which scales the inputs and then produces an output, similar to the axon of the neuron.

More succinctly, the inputs fed into the network are scaled based on the applied weights, resulting in an output that is summarized across all those weighted inputs. Conceptually this is not unlike an algorithmic formula wherein assigned variables are applied to produce an output measure. However, algorithms do not typically have variable “weights” that can be altered based on a goal (called “training”), which makes the ANN a significantly different application.

Given that one can see the network infrastructure based on the descriptions above, the next component to address is how those weights get adjusted (that goal-oriented training just mentioned). While there are a large number of algorithms available for applying such adjustments (as Grajczyk, 2008, p. 10, states, “For every type of ANN there are several learning algorithms”), the key is that these adjustments are not simply applied but are learned. The mechanism used for ANN learning is described by the Widrow-Hoff Learning Rule, which Abdi, Valentin, and Edelman (1999, p. 9) summarize as, “...when you make a mistake, pay less attention to the input cells that told you to make this mistake, and pay more attention to the input cells that told you not to make this mistake.” These ‘attention’ adjustments are effectively the weights applied to the inputs – change them and they change the effect the input has on the next perceptron and, hence, on the network output.

By far the most common learning (weight-adjustment) algorithm in use by ANNs is called back-propagation (Garson, 1998, p. 33; Stergiou & Siganos, 1997, Appendix A), or as

Caudill and Butler (1990) suggest, “...backward error propagation learning” (p. 183). As Garson describes it, back-propagation, “...modifies input weights on the basis of error signals arising from the output layer” (1998, p. 42). That is, much as we humans learn by examining what we did wrong, and then make behavioral adjustments, the ANN, given a known set of Inputs and Outputs (a *gold standard* from which to learn), makes adjustments to its weights based on the variance (error) from the Output that the ANN projected given those Inputs. Hence, this is why Caudill and Butler explicitly refer to back-propagation as a learning algorithm and not a network design (1990, p. 184).

Validation and Testing

One additional issue to examine is how the ANN’s output is validated. The ANN network is configured from the data used the learning algorithm specified, and based on its output’s success (as compared with the known *gold standard* which it was evaluated against during the learning process), it achieves some predictive measure that represents how well it learned. One such predictive measure is a *performance rate*, which indicates, as a percentage, how many of the training cases did its outcome match (that is, how many cases did it classify correctly). Its correctness can be easily determined in the context of the training set (the data used to train the network), but that does not address how well the network would perform across the full population – e.g., how generalizable the network is. For that one must apply a different set of data from the training set to see if the outcomes of the network are reasonably close to those it was trained with. If the training set was highly representative of the full population, then the likelihood that it will be generalizable is fairly strong. However, if the training set were not representative of the population then the network is considered to be *overfitted* (Bartosch-Härlid et al., 2008; Lancashire et al., 2009).

This process is often referred to as ANN validation, and it is typically accomplished by either holding back a segment of the training data (that is, the training data is a randomly selected portion of all of the data gathered, while the remainder of that dataset becomes the validation set), or separate data is collected and examined post-training. Sometimes there is a third set created specifically for testing – after training and after validation, a third collection of data is examined to determine the network’s true performance. Usually, however, there are only two runs used – testing and validation (sometimes *testing* and *validation* are used interchangeably).

Additional Resources

Thus, ANNs present us with a tool based on very recent technologies that allow us to give “voice” to large databases (data sets) given the specific constraints of output (which, in healthcare, we call “outcome”) and a large collection of inputs (the wide array of clinical detail that can be associated with the clinical outcome selected). However, for those interested in pursuing further information on the technologies and algorithms employed by ANN tools, the following readings, presented in alphabetical order, are suggested (noting that while these are all of 1990s vintage, that timeframe coincides with the beginning of current ANN development and application, hence their context is still meaningful):

- Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Thousand Oaks, CA: SAGE Publications, Inc.
- Bigus, J. P. (1996). *Data mining with neural networks*. New York, NY: McGraw-Hill
- Caudill, M., & Butler C. (1990). *Naturally intelligent systems*. Cambridge, MA: Massachusetts Institute of Technology.
- Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. Thousand Oaks, CA: Sage Publications Ltd.

The Value of ANNs to Healthcare Research

Next is an examination of why ANNs are used in research, and how prominent they are in recent healthcare studies. In order to answer the first part it should be recognized that ANNs are, by definition, evidence-based – they are machine-learning systems which derive their logic solely from the data presented to them. As Rousseau (2006, p. 258) pointed out, evidence-based clinical care as a way of life in healthcare is of relatively recent vintage, with its greatest growth occurring after 1990. And while the key elemental building block of ANNs, the *perceptron*, was first suggested by McCulloch and Pitts back in 1943 (Caudill & Butler, 1990, p.163; Lancashire et al., 2009, p. 316), it was not until the 1990s that digital computing power became cheap and fast enough to warrant the development of ANN applications (Garson, 1998, p. 16). Next, the evidence for their increased use is provided graphically in Appendix D, which shows two independently derived graphs of ANN-based research publications and clinical trial usage during the 1990s (Gant, Rodway, & Wyatt, 2001) and early 2000s (Lisboa & Taktak, 2006).

These charts clearly indicate an upward trend during their respective timeframes, and it is suggested here that this is not coincidental to the expansion of evidence-based medicine. Indeed, Barends, ten Have, and Huisman (2012) indicated the shift in medicine to evidence-based approaches has been causing a change in emphasis in the kinds of research questions being dealt with in the literature. They posited that background questions, about general knowledge of the biomedical aspects of a disease or disorder, were no longer the primary focus of research, while foreground questions, those which are about specific knowledge that can be used in clinical decision-making about the treatment of a patient, were gaining favor (Barends, ten Have, & Huisman, 2012, p. 33). It is exactly those foreground evidence-based questions for which ANNs are most suitably designed.

Thus we have not only new technology being employed in the realm of medical research, but one that is doing so based principally on clinical evidence. As Grajczyk (2008) suggested in a paper from a seminar on decision support systems, an ANN helps to solve complex problems through learning (p. 10). Unlike expert systems (those which are rule-based, employing a set of human-expert derived instructions in a pre-defined process to solve a problem), the ANN examines the only the data presented, without direct human intervention. Using methodologies such as supervised learning through backpropagation (the most common methodological variety), the network iteratively derives the relationships between these data, called Inputs, and their outcomes, or Outputs (Caudill & Butler, 1990, pp. 183-184). Lancashire et al. (2008, p. 318) further posit that once ANNs are trained in this way, they present a real-world solution to a given problem by their ability to predict future cases or trends based on the data from which they were created. It is here that ANN-based analysis is differentiated from other technology-based decision-automation systems, and how their generalist capabilities are similar to the neural networks of biological systems. Indeed, Lancashire et al. (2009, p. 316) contend that, “ANNs are inspired by the way in which the human brain learns and processes information...to generalise and predict well for future cases.”

Another aspect of ANN technology is its ability to parse vastly large databases for difficult to recognize relationships. The first part of that requires the provision of large-scale database-driven applications that have a strong degree of standardization for coding and content. This very process has already begun within the healthcare industry, and is in no small part due to the implementation by the U.S. Government of the Health Information Technology for Economic and Clinical Health (HITECH) Act, a part of the American Reinvestment and Recovery Act of 2009 (2009), or ARRA. This Act provided substantial stimulus funding towards advancing

electronic health record (EHR) technology implementations (Litwin, 2011, p. 864), enabling the foundation for large and relatively standardized databases. The realization of this becomes clear when one examines that, according to the National Institutes of Health (NIH) online public health reports (Burke, 2010, p. 141) the U.S. Congress allocated over \$49 billion in federal spending in support of EHR implementation. Thus, the seeds may have already been sown for an emerging evidentiary data source from which ANN applications can ultimately pool analyses.

Appendix B

The PRISMA Statement

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	6
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	10
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	24
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	n/a
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	63
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	64
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	65
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	67
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	74

Section/topic	#	Checklist item	Reported on page #
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	75
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	70
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	73
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	80
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	80 , 92
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	84
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	74
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Chapter 5
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	73
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Appendix I
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	95
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	107
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	85

Section/topic	#	Checklist item	Reported on page #
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	80
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	107
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Chapter 6
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	n/a

Note: Adapted from Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097. Those showing "n/a" under the "Reported" column reflect that the related Section/Topic was deemed not relevant to this particular study.

Appendix C

Predictive Tool Comparison

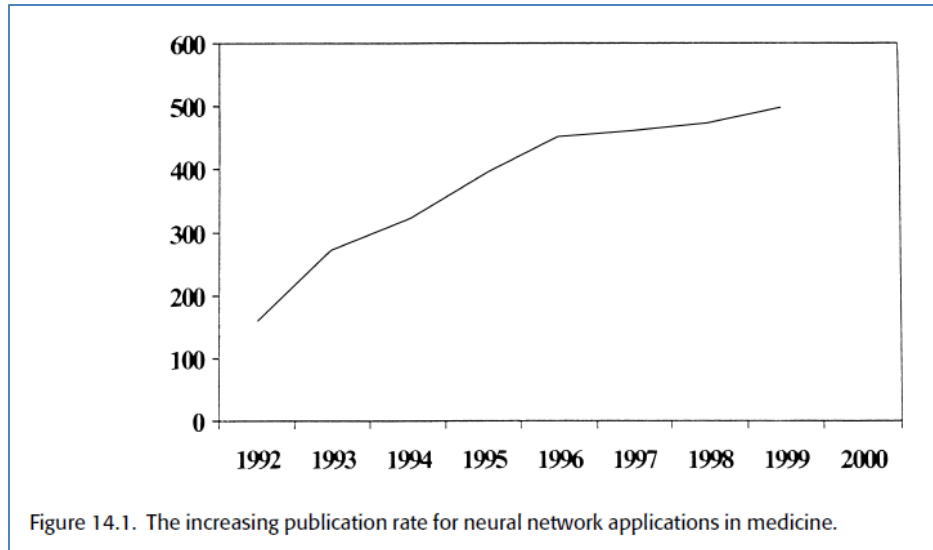
Criterion	Specific requirement	Artificial neural nets	Explicit statistical models	Knowledge-based systems
1. Accuracy	Accurate discrimination	✓	✓	?
	Well-calibrated probabilities	×	✓?	×
2. Generality	Valid when transferred to other sites	?	?	?
	Model can be adjusted to reduce overoptimism	×	✓	?
3. Clinical credibility	Model's structure apparent, explanations available	×✓?	✓	
	Ability to browse the system's 'knowledge'	×	✓	✓
	Simple to calculate predictions	×	✓?	×
	Ability to display 'common sense'	×	×	?
4. Ease of development	Avoids need for large, prospective verified database	×	×	✓?
	Avoids need for skilled personnel	?	?	×
	Ability to encode clinical policy, systematic review results, etc.	×	×	✓
	Ability to encode aetiology, disease mechanisms, etc.	×	✓	✓
	Ability to learn from experience	✓	✓	×
5. Clinical effectiveness	RCT evidence of impact on clinical process, patient outcome ^a	×	✓	✓

From Wyatt & Altman 1995.
^aRCT, randomized clinical trials; data from Hunt et al. 1998.

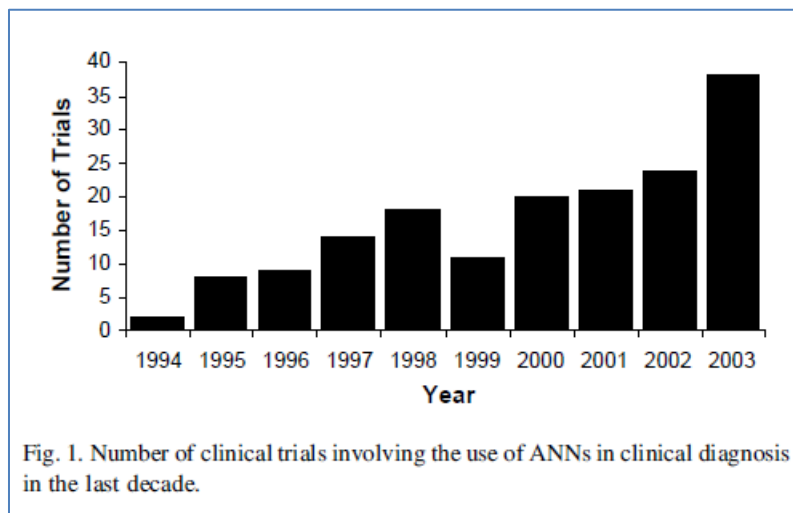
Note: Table indicating criteria for predictive tools and comparison of the ability of three predictive technologies to satisfy them. Adapted from Gant, Rodway, & Wyatt (2001, p. 352, Table 14.3). Copyright 2001 by Cambridge University Press. × = No; ✓ = Yes; ? = Unknown; ×✓? = Unsure.

Appendix D

Artificial Neural Network (ANN) Application Usage in Healthcare Research Publications



Adapted from Gant, Rodway, & Wyatt, 2001, p. 330. Copyright 2001 by Cambridge University Press



Adapted from Lisboa & Taktak, 2006, p.2090. Copyright 2006 by Elsevier.

Appendix E

Proposition Logic Chain

Proposition	Theoretical Basis	Reference(s)
1 Humans are to be flawed decision makers.	Clinician decision makers apply natural human bias to clinical decisions.	Croskerry, 2002; Croskerry, 2014; Ferreira et al., 2010; Mendel et al., 2011
	Biased clinical decisions result in inappropriate, and costly (in terms of both clinical and financial outcomes), patient services.	Berner & Graber, 2008; Croskerry, 2002; Graber, Franklin, & Gordon, 2005; Institute of Medicine, 1999; Pham et al., 2012
2 Established cognitive mechanisms contribute to human analytical limitations.	Satisficing - reaching a <i>good enough</i> decision based on the discretionary judgment of the clinician, but that may truncate the complete analysis needed to determine the correct decision.	Simon, 1943; Simon, 1979; Simon, 1997
	Sensemaking - in trying to understand an emerging reality in a crisis condition (such as a major trauma even in an ED), decision accuracy can be trumped by decision certainty (confidence) as a means to reduce external pressures and stress.	Weick, 2005; Weick, Sutcliffe, & Obstfield, 2005
	Heuristic thinking - <i>rules of thumb</i> applied to decisions, while fast-tracking the decision process, can cause the decision-maker to overlook critical stages of the process, resulting in inappropriate decisions.	Kahneman, 2011; Tversky & Kahneman, 1974; West, Toplak, & Stanovich, 2008

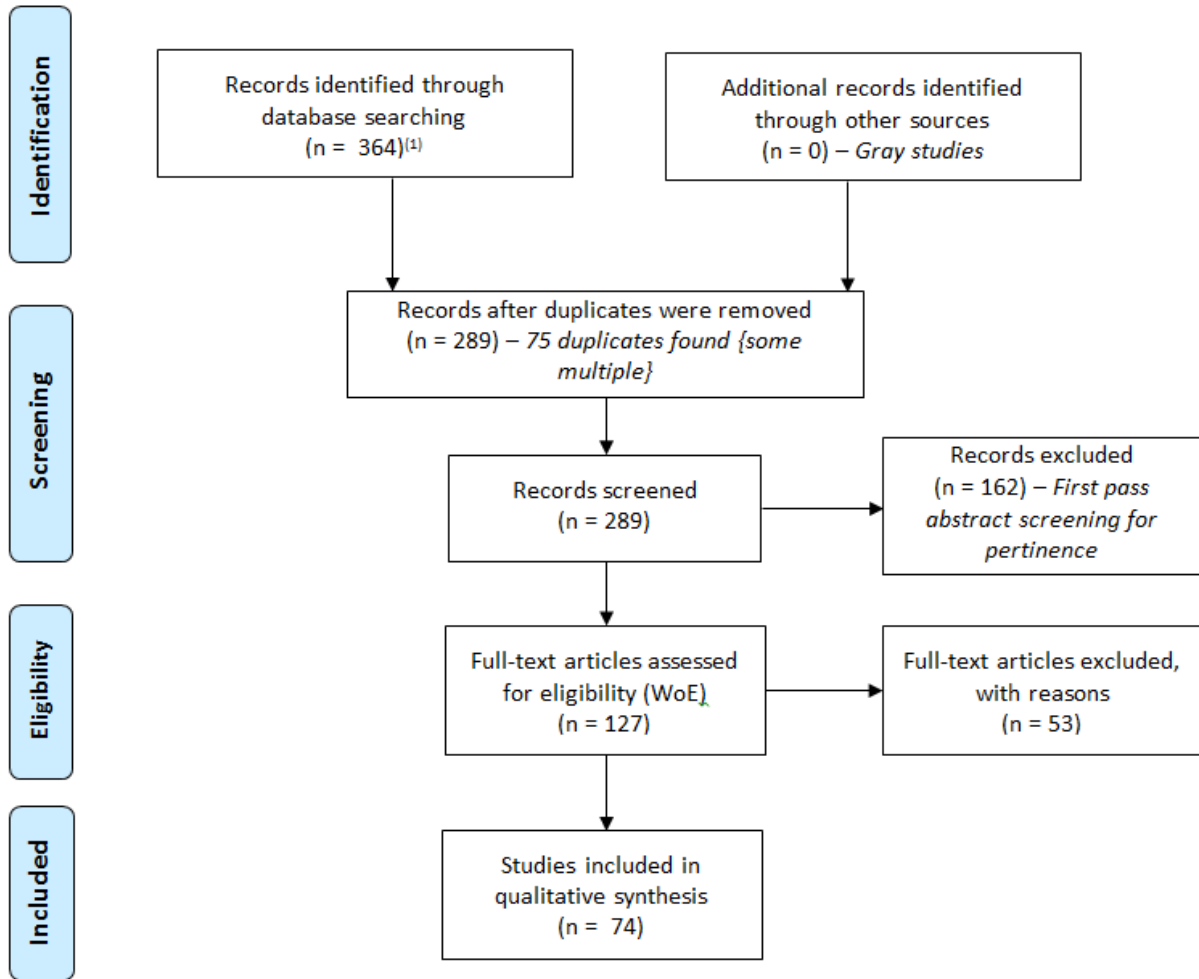
Proposition	Theoretical Basis	Reference(s)
3 ANNs provide evidence-based decision networks, given sufficient data and applied training, which have been commonly applied in research.	A shift has been noted in the literature from general knowledge questions (background) to treatment questions (foreground) of the type that an ANN would be used to address.	Barends, ten Have, & Huisman, 2012)
	Data collection in healthcare, as needed for ANN creation, has been greatly augmented by the implementation of EMR systems, aided by regulatory-driven funding.	Litwin, 2011; Recovery.gov, 2014
	ANNs are machine learning tools that are trained on collected datasets which are determined to be pertinent to the condition/malady being assessed by the clinician.	Caudill & Butler, 1990; Garson, 1998, Lancashire et al., 2009
	Automated clinical decision support tools, such as ANN technologies, have been studied and found to provide both risk-aversion and error mitigation properties.	Croskerry, 2002; Ferreira et al., 2010; Jaspers et al, 2011
4 ANNs have been almost exclusively engaged in research but only rarely applied to practice.	Very few examples exist of ANNs' being used to directly inform patient care decisions.	Gant et al., 2001
	Criteria for determination of how <i>effective</i> ANNs are (i.e., their predictive validity) have been established.	Collopy, Adya, & Armstrong, 1994
5 ANN use in research has shown them to be effective decision-making tools.	<i>Research study applied within the current paper.</i>	
6 Employing ANNs as a <i>nudge</i> to the clinical decision process might prove effective toward improving decision outcome.	The "Data Refinery" model proposed by Gant et al. (2001) which suggests how ANNs might be incorporated into the clinical decision process.	Gant et al., 2001

Proposition	Theoretical Basis	Reference(s)
6 (continued)	When DRAWN is engaged by the clinician, the ANN could offset the negative effects of cognitive biases and heuristics by giving either a confirmatory or contradictory <i>nudge</i> .	Thaler, 2000; Thaler & Sunstein, 2009
7	The cost of ANN implementation can be offset by the resultant reduction in clinical judgment error.	Institute of Medicine, 1999; Pham et al., 2012
	Costs associated with errors in clinical decision-making are well established.	
	Costs associated with technology are typically evaluated by risk reductions rather than increased revenue generation.	Whalen, 2015

Appendix F



PRISMA 2009 Flow Diagram



(1) Databases Accessed:

- UMUC Library OneSearch - 64
- iCONN (Univ. of Conn. OneSearch) – 100 most relevant of 11,000
- NLM/NIH PubMed – 200 most relevant of 17,000

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit www.prisma-statement.org.

Appendix G

References for Analyzed Studies

- Abouzari, M., Rashidi, A., Zandi-Toghiani, M., Behzadi, M., & Asadollahi, M. (2009). Chronic subdural hematoma outcome prediction using logistic regression and an artificial neural network. *Neurosurgical Review, 32*(4), 479-484. doi:10.1007/s10143-009-0215-3
- Acton, P. D., & Newberg, A. (2006). Artificial neural network classifier for the diagnosis of Parkinson's disease using [99mTc]TRODAT-1 and SPECT. *Physics in Medicine and Biology, 51*(12), 3057-3066. doi:10.1088/0031-9155/51/12/004
- Aggarwal, Y., Karan, B., Das, B., Aggarwal, T., & Sinha, R. (2007). Backpropagation ANN-Based prediction of exertional heat illness. *Journal of Medical Systems, 31*(6), 547-550. doi: 10.1007/s10916-007-9097-5
- Ahmad, F., Mat Isa, N., Hussain, Z., & Sulaiman, S. (2013). A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis. *Neural Computing & Applications, 23*(5), 1427-1435. doi:10.1007/s00521-012-1092-1
- Alkan, A., Koklukaya, E., & Subasi, A. (2005). Automatic seizure detection in EEG using logistic regression and artificial neural network. *Journal of Neuroscience Methods, 148*(2), 167-176. doi:10.1016/j.jneumeth.2005.04.009
- Amodio, P., Pellegrini, A., Ubiali, E., Mathy, I., Piccolo, F. D., Orsato, R., & ... Guerit, J. (2006). The EEG assessment of low-grade hepatic encephalopathy: Comparison of an artificial neural network-expert system (ANNES) based evaluation with visual EEG readings and EEG spectral analysis. *Clinical Neurophysiology, 117*(10), 2243-2251. doi:10.1016/j.clinph.2006.06.714

- Andersson, S., Heijl, A., Bizios, D., & Bengtsson, B. (2013). Comparison of clinicians and an artificial neural network regarding accuracy and certainty in performance of visual field assessment for the diagnosis of glaucoma. *Acta Ophthalmologica*, 91(5), 413-417. doi:10.1111/j.1755-3768.2012.02435.x
- Aparicio Castillo, A., De La Rosa Vázquez, J. M., Calva Chavarría, P. A., Franco López, E. B., Torres Manzo, R., Álvarez Dorantes, R.,...Romero Guadarrama, M. B. (2006). Evaluation of healthy and infected cervical tissue using a LIFS system and a back-propagation neural network. *Revista Mexicana de Ingeniería Biomédica*, 27(2), 68-73. Retrieved from <http://new.medigraphic.com/>
- Avci, D., Leblebicioglu, M., Poyraz, M., & Dogantekin, E. (2014). A New Method Based on Adaptive Discrete Wavelet Entropy Energy and Neural Network Classifier (ADWEENN) for Recognition of Urine Cells from Microscopic Images Independent of Rotation and Scaling. *Journal of Medical Systems*, 38(2), 1-9. doi:10.1007/s10916-014-0007-3
- Aydin, S., Saraoğlu, H. M., & Kara, S. (2011). Singular spectrum analysis of sleep EEG in insomnia. *Journal of Medical Systems*, 35(4), 457-461. doi:10.1007/s10916-009-9381-7
- Behrman, M., Linder, R., Assadi, A. H., Stacey, B. R. and Backonja, M. (2006), Classification of patients with pain based on neuropathic pain symptoms: Comparison of an artificial neural network against an established scoring system. *European Journal of Pain*, 11(4), 370–376. doi:10.1016/j.ejpain.2006.03.001
- Bevevino, A., Dickens, J., Potter, B., Dworak, T., Gordon, W., & Forsberg, J. (2014). A model to predict limb salvage in severe combat-related open calcaneus fractures. *Clinical Orthopaedics and Related Research*, 472(10), 3002-3009. doi:10.1007/s11999-013-3382-z

- Bhatikar, S. J., DeGross, C., & Mahajan, R. L. (2005). A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics. *Artificial Intelligence in Medicine*, 33(3), 251-260. doi:10.1016/j.artmed.2004.07.008
- Biglarian, A., Bakhshi, E., Gohari, M., & Khodabakhshi, R. (2012). Artificial neural network for prediction of distant metastasis in colorectal cancer. *Asian Pacific Journal of Cancer Prevention*, 13(3), 927-930. doi:10.7314/APJCP.2012.13.3.927
- Bollschweiler, E. H., Monig, S. P., Hensler, K., Baldus, S. E., Maruyama, K., & Holscher, A. H. (2004). Artificial neural network for prediction of lymph node metastases in gastric cancer: A phase II diagnostic study. *Annals of Surgical Oncology*, 11(5), 506–511. doi:10.1245/aso.2004.04.018
- Carrara, M., Bono, A., Bartoli, C., Colombo, A., Lualdi, M. Moglia, D.,..., & Marchesini, R. (2007). Multispectral imaging and artificial neural network: Mimicking the management decision of the clinician facing pigmented skin lesions. *Physics in Medicine and Biology*, 52(9), 2599-2613. doi:10.1088/0031-9155/52/9/018
- Chen, M., & Chou, C. (2014). Applying Cybernetic Technology to Diagnose Human Pulmonary Sounds. *Journal of Medical Systems*, 38(6), 1-10. doi:10.1007/s10916-014-0058-5
- Cheng, B., Joe Stanley, R., Stoecker, W. V., Stricklin, S. M., Hinton, K. A., Nguyen, T. K.,...Moss, R. (2013). Analysis of clinical and dermoscopic features for basal cell carcinoma neural network classification. *Skin Research and Technology*, 19(1), e217-e222. doi:10.1111/j.1600-0846.2012.00630.x
- Chiu, J., Wang, Y., Su, Y., Wei, L., Liao, J., & Li, Y. (2009). Artificial neural network to predict skeletal metastasis in patients with prostate cancer. *Journal of Medical Systems*, 33(2), 91-100. doi: 10.1007/s10916-008-9168-2

- Cho, K., Müller, J. H., Scheffer, C., & Erasmus, P. J. (2013). Application of an artificial neural network for the quantitative classification of trochlear dysplasia. *Journal of Mechanics in Medicine & Biology*, 13(4), 1350059-1-1350059-14. doi:10.1142/S0219519413500590
- Chun, F., Briganti, A., Gallina, A., Karakiewicz, P., Hopp, J., Huland, H., & ... Kattan, M. (2007). Initial biopsy outcome prediction-head-to-head comparison of a logistic regression-based nomogram versus artificial neural network. *European Urology*, 51(5), 1236-1243. doi:10.1016/j.eururo.2006.07.021
- Delavarian, M., Towhidkhalah, F., Dibajnia, P., & Gharibzadeh, S. (2012). Designing a decision support system for distinguishing ADHD from similar children behavioral disorders. *Journal of Medical Systems*, 36(3), 1335-1343. doi:10.1007/s10916-010-9594-9
- Dey, P., Logasundaram, R., & Joshi, K. (2013). Artificial neural network in diagnosis of lobular carcinoma of breast in fine-needle aspiration cytology. *Diagnostic Cytopathology*, 41(2), 102-106. doi:10.1002/dc.21773
- Diao, X., Zhang, X., Wang, T., Chen, S., Yang, Y., & Zhong, L. (2011). Highly sensitive computer aided diagnosis system for breast tumor based on color doppler flow images. *Journal of Medical Systems*, 35(5), 801-809. doi:10.1007/s10916-010-9461-8
- Dietzel, M., Baltzer, P. A. T., Dietzel, A., Zoubi, R., Gröschel, T., Burmeister, H. P., . . . Kaiser, W. A. (2012). Artificial Neural Networks for differential diagnosis of breast lesions in MR-Mammography: A systematic approach addressing the influence of network architecture on diagnostic performance using a large clinical database. *European Journal of Radiology*, 81(7), 1508-1513. doi:10.1016/j.ejrad.2011.03.024

- Erol, R., Ogulata, S., Sahin, C., & Alparslan, Z. (2008). A Radial Basis Function Neural Network (RBFNN) approach for structural classification of thyroid diseases. *Journal of Medical Systems*, 32(3), 215-220. doi: 10.1007/s10916-007-9125-5
- Farhadian, M., Aliabadi, M., & Darvishi, E. (2015). Empirical estimation of the grades of hearing impairment among industrial workers based on new artificial neural networks and classical regression methods. *Indian Journal of Occupational & Environmental Medicine*, 19(2), 84-89. doi:10.4103/0019-5278.165337
- Fernández, E. A., Valtuille, R., Presedo, J. R., & Willshaw, P. (2005). Comparison of standard and artificial neural network estimators of hemodialysis adequacy. *Functional Ecology*, 29(2), 159-165. doi:10.1111/j.1525-1594.2005.29027.x
- Fraschini, M. (2011). Mammographic masses classification: novel and simple signal analysis method. *Electronics Letters*, 47(1), 14-15. doi: 10.1049/el.2010.2712
- Gargouri, N., Dammak Masmoudi, A., Sellami Masmoudi, D., & Abid, R. (2012). A New GLLD Operator for Mass Detection in Digital Mammograms. *International Journal of Biomedical Imaging*, 2012, 1-13. doi:10.1155/2012/765649
- Geng, Z., Hoffman, M. R., Jones, C. A., McCulloch, T. M., & Jiang, J. J. (2013). Three-dimensional analysis of pharyngeal high-resolution manometry data. *Laryngoscope*, 123(7), 1746-1753. doi:10.1002/lary.23987
- Gheonea, D., Streba, C., Vere, C., Rogoveanu, I., Șerbănescu, M., Pirici, D., & ... Mogoantă, S. (2014). Diagnosis system for hepatocellular carcinoma based on fractal dimension of morphometric elements integrated in an artificial neural network. *Biomed Research International*, 2014, 1-10. doi:10.1155/2014/239706

- Halford, J. J., Schalkoff, R. J., Zhou, J., Benbadis, S. R., Tatum, W. O., Turner, R. P., & ... Dean, B. C. (2013). Standardized database development for EEG epileptiform transient detection: EEGnet scoring system and machine learning analysis. *Journal of Neuroscience Methods*, 212(2), 308-316. doi:10.1016/j.jneumeth.2012.11.005
- Hallner, D., & Hasenbring, M. (2004). Classification of psychosocial risk factors (yellow flags) for the development of chronic low back and leg pain using artificial neural network. *Neuroscience Letters*, 361(1-3), 151-154. doi:10.1016/j.neulet.2003.12.107
- He, X., Sahiner, B., Gallas, B., Chen, W., & Petrick, N. (2014). Computerized characterization of lung nodule subtlety using thoracic CT images. *Physics in Medicine & Biology*, 59(4), 897-910. doi:10.1088/0031-9155/59/4/897
- Hoffman, M., Jones, C., Geng, Z., Abelhalim, S., Walczak, C., Mitchell, A., . . . McCulloch, T. (2013a). Classification of high-resolution manometry data according to videofluoroscopic parameters using pattern recognition. *Otolaryngology - Head and Neck Surgery*, 149(1), 126-133. doi:10.1177/0194599813489506
- Hoffman, M. R., Mielens, J. D., Omari, T. I., Rommel, N., Jiang, J. J., & McCulloch, T. M. (2013b). Artificial neural network classification of pharyngeal high-resolution manometry with impedance data. *The Laryngoscope*, 123(3), 713–720. doi:10.1002/lary.23655
- Hossen, A. (2013). A neural network approach for feature extraction and discrimination between Parkinsonian tremor and essential tremor. *Technology & Health Care*, 21(4), 345-356. doi:10.3233/THC-130735
- Huang, M. L., Hung, Y. H., Lee, W. M., Li, R. K., & Wang, T. H. (2012). Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification

- techniques in breast cancer dataset classification diagnosis. *Journal of Medical Systems*, 36(2), 407-414. doi:10.1007/s10916-010-9485-0
- Hui, E. P., Leung, L. K. S., Poon, T. C. W., Mo, F., Chan, V. T. C., Ma, A. T. W., . . . Chan, A. T. C.. (2011). Prediction of outcome in cancer patients with febrile neutropenia: A prospective validation of the Multinational Association for Supportive Care in Cancer risk index in a Chinese population and comparison with the Talcott model and artificial neural network. *Supportive Care in Cancer*, 19(10), 1625-1635. doi:10.1007/s00520-010-0993-8
- Ibrahim, F., Faisal, T., Mohamad Salim, M. I., Taib, M. N. (2010). Non-invasive diagnosis of risk in dengue patients using impedance analysis and artificial neural network. *Medical & Biological Engineering & Computing*, 48(11), 1141-1148. doi:10.1007/s11517-010-0669-z
- Işik, H. & Arslan, S. (2011). An Artificial Neural Network classification approach for use the ultrasound in physiotherapy. *Journal of Medical Systems*, 35(6), 1333-1341. doi:10.1007/s10916-009-9410-6
- Joo, S., Yang, Y. S., Moon, W. K., & Kim, H. C. (2004). Computer-aided diagnosis of solid breast nodules: Use of an artificial neural network based on multiple sonographic features. *IEEE Transactions on Medical Imaging*, 23(10), 1292-1300. doi:10.1109/iembs.2004.1403434
- Jovanovic, P., Salkic, N., & Zerem, E. (2014). Artificial neural network predicts the need for therapeutic ERCP in patients with suspected choledocholithiasis. *Gastrointestinal Endoscopy*, 80(2), 260-268. doi:10.1016/j.gie.2014.01.023

Kara, S., & Güven, A. (2007). Neural network-based diagnosing for optic nerve disease from visual-evoked potential. *Journal of Medical Systems, 31*(5), 391-396. doi:

10.1007/s10916-007-9081-0

Kaya, Y. (2014). A fast intelligent diagnosis system for thyroid diseases based in extreme learning machine. *Anadolu University of Sciences & Technology - A: Applied Sciences & Engineering, 15*(1), 41-49. doi:10.18038/btd-a.89202

Kewk, L. C., Fu, S., Chia, T. C., Diong, C. H., Tang, C. L., & Krishnan, S. M. (2005). High-sensitivity and specificity of laser-induced autofluorescence spectra for detection of colorectal cancer with an artificial neural network. *Applied Optics, 44*(19), 4004-4008. doi:10.1364/AO.44.004004

Koçer, S. & Canal, M.R. (2009). Classifying epilepsy diseases using artificial neural networks and genetic algorithm. *Journal of Medical Systems, 35*(4), 489-498. doi:10.1007/s10916-009-9385-3

Kocyigit, Y., Alkan, A., & Erol, H. (2008). Classification of EEG recordings by using Fast Independent Component Analysis and Artificial Neural Network. *Journal of Medical Systems, 32*(1), 17-20. doi: 10.1007/s10916-007-9102-z

Kshirsagar, A., Seftel, A., Ross, L., Mohamed, M., & Niederberger, C. (2006). Predicting hypogonadism in men based upon age, presence of erectile dysfunction, and depression. *International Journal of Impotence Research, 18*(1), 47-51. doi:10.1038/sj.ijir.3901369

Kumar, Y., and Sahoo, G. (2013). Prediction of different types of liver diseases using rule based classification model. *Technology and Health Care, 21*(5), 417-432. doi:10.3233/THC-130742

- Kuruvilla, J., & Gunavathi, K. (2014). Lung cancer classification using neural networks for CT images. *Computer Methods and Programs in Biomedicine*, 113(1), 202-209.
doi:10.1016/j.cmpb.2013.10.011
- La Torre, E., Caputo, B., & Tommasi, T. (2010). Learning Methods for Melanoma Recognition. *International Journal of Imaging Systems and Technology*, 20(4), 316-322. doi: 10.1002/ima.20261
- Lai, M., De Stefano, V., & Landolfi, R. (2014). Autoimmune hemolytic anemia with gel-based immunohematology tests: Neural network analysis. *Immunologic Research*, 58(1), 70-74.
doi:10.1007/s12026-013-8480-1
- Li, L., Zhang, Q., Ding, Y., Jiang, H., Thiers, B. H., & Wang, J. Z. (2014). Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system. *BMC Medical Imaging*, 14(1), 1-12. doi:10.1186/1471-2342-14-36
- Limonadi, F. M., McCartney, S., & Burchiel, K. J. (2006). Design of an artificial neural network for diagnosis of facial pain syndromes. *Stereotactic and Functional Neurosurgery*, 84(5-6), 212-220. doi:10.1159/000095167
- Lin, H., (2008). Identification of spinal deformity classification with total curvature analysis and artificial neural network. *IEEE Transactions on Biomedical Engineering*, 55(1), 376-382.
doi:10.1109/TBME.2007.894831
- Linder, R., Albers, A. E., Hess, M., Pöpl, S. J., & Schönweiler, R. (2008). Artificial neural network-based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. *Journal of Voice*, 22(2), 155-163.
doi:10.1016/j.jvoice.2006.09.003

- Liu, H., Tang, Z., Yang, Y., Weng, D., Sun, G., Duan, Z., & Chen, J. (2009). Identification and classification of high risk groups for Coal Workers' Pneumoconiosis using an artificial neural network based on occupational histories: a retrospective cohort study. *BMC Public Health*, 9(366), 1-8. doi: 10.1186/1471-2458-9-366
- Lux, A., Müller, R., Tulk, M., Olivieri, C., Zarrabeita, R., Salonikios, T., & Wirnitzer, B. (2013). HHT diagnosis by mid-infrared spectroscopy and artificial neural network analysis. *Orphanet Journal of Rare Diseases*, 8(1), 94-108. doi:10.1186/1750-1172-8-94
- Lweesy, K., Fraiwan, L., Khasawneh, N., & Dickhaus, H. (2011). New automated detection method of OSA based on artificial neural networks using P-Wave shape and time changes. *Journal of Medical Systems*, 35(4), 723-734. doi:10.1007/s10916-009-9409-z
- Mariani, S., Manfredini, E., Rosso, V., Grassi, A., Mendez, M., Alba, A., & ... Bianchi, A. (2012). Efficient automatic classifiers for the detection of A phases of the cyclic alternating pattern in sleep. *Medical & Biological Engineering & Computing*, 50(4), 359-372. doi:10.1007/s11517-012-0881-0
- Matulewicz, L., Jansen, J. F.A., Bokacheva, L., Vargas, H. A., Akin, O., Fine, S. W., . . . & Zakian, K. L. (2014), Anatomic segmentation improves prostate cancer detection with artificial neural networks analysis of 1H magnetic resonance spectroscopic imaging. *Journal of Magnetic Resonance Imaging*, 40(6),1414–1421. doi:10.1002/jmri.24487
- McLaren, C. E., Chen, W., Nie, K., & Su, M. (2009). Prediction of malignant breast lesions from MRI features: a comparison of artificial neural network and logistic regression techniques. *Academic Radiology*, 16(7), 842–851. doi:10.1016/j.acra.2009.01.029

- Mei-Ling, H., Hsin-Yi, C., & Por-Tying, H. (2006). Analysis of Glaucoma Diagnosis with Automated Classifiers using Stratus Optical Coherence Tomography. *Optical & Quantum Electronics*, 37(13-15), 1239-1249. doi:10.1007/s11082-005-4195-4
- Mert, A., Kilic, N., Bilgili, E., & Akan, A. (2015). Breast cancer detection with reduced feature set. *Computational and Mathematical Methods in Medicine*, 1-11. doi:10.1155/2015/265138
- Michalski, R., Wit, A., Gajewski, J. (2011). Use of artificial neural networks for assessing parameters for gait symmetry. *Acta of Bioengineering and Biomechanics*, 13(4), 65-70. Retrieved from <http://www.actabio.pwr.wroc.pl/>
- Mielens, J. D., Hoffman, M. R., Ciucci, M. R., McCulloch, T. M., & Jianga, J. J. (2012). Application of Classification Models to Pharyngeal High-Resolution Manometry. *Journal of Speech, Language & Hearing Research*, 55(3), 892-902. doi:10.1044/1092-4388(2011/11-0088)
- Mofidi, R., Deans, C., Duff, M. D., de Beaux, A. C., & Paterson Brown, S. (2006). Prediction of survival from carcinoma of oesophagus and oesophago-gastric junction following surgical resection using an artificial neural network. *European Journal of Surgical Oncology*, 32(5), 533-539. doi:10.1016/j.ejso.2006.02.020
- Mohamed, H., Mabrouk, M. S., & Sharawy, A. (2014). Computer aided detection system for micro calcifications in digital mammograms. *Computer Methods and Programs in Biomedicine*, 116(3), 226-235. doi:10.1016/j.cmpb.2014.04.010
- Nayak, G., Kamath, S., Pai, K., Sarkar, A., Ray, S., Kurien, J., . . . Mahato, K. (2006). Principal component analysis and artificial neural network analysis of oral tissue fluorescence

- spectra: Classification of normal premalignant and malignant pathological conditions. *Biopolymers*, 82(2), 152-166. doi:10.1002/bip.20473
- Niwas, S. I., Palanisamy, P., Chibbar, R., & Zhang, W. (2012). An Expert Support System for Breast Cancer Diagnosis using Color Wavelet Features. *Journal of Medical Systems*, 36(5), 3091-3102. doi:10.1007/s10916-011-9788-9
- Norman, R. G., Rapoport, D. M., & Ayappa, I. (2007). Detection of flow limitation in obstructive sleep apnea with an artificial neural network. *Physiological Measurement*, 28(9), 1089-1100. doi:10.1088/0967-3334/28/9/010
- Olsson, S., Ohlsson, M., Ohlin, H., Dzaferagic, S., Nilsson, M., Sandkull, P., & Edenbrandt, L. (2006). Decision support for the initial triage of patients with acute coronary syndromes. *Clinical Physiology and Functional Imaging*, 26(3), 151-156. doi: 10.1111/j.1475-097x.2006.00669.x
- Özbay, Y. (2009). A new approach to detection of ECG arrhythmias: complex discrete wavelet transform based complex valued artificial neural network. *Journal of Medical Systems*, 33(6), 435-445. doi: 10.1007/s10916-008-9205-1
- Pachori, R. B., & Patidar, S. (2014). Epileptic seizure classification in EEG signals using second-order difference plot of intrinsic mode functions. *Computer Methods and Programs in Biomedicine*, 113(2), 494-502. doi:10.1016/j.cmpb.2013.11.014
- Park, S. C., Tan, J., Wang, X., Lederman, D., Leader, J. K., Kim, S. H., & Zheng, B. (2011). Computer-aided detection of early interstitial lung diseases using low-dose CT images. *Physics in Medicine and Biology*, 56(4), 1139–1153. doi:10.1088/0031-9155/56/4/016
- Pellegrini, A., Ubiali, E., Orsato, R., Schiff, S., Gatta, A., Castellaro, A.,..., & Amodio, P. (2005). Electroencephalographic staging of hepatic encephalopathy by an artificial neural

- network and an expert system. *Neurophysiologie Clinique*, 35(5-6), 162-167.
doi:10.1016/j.neucli.2005.12.003
- Peng, S. Y., Wu, K. C., Wang, J. J., Chuang, J. H., Peng, S. K., & Lai, Y. H. (2006). Predicting postoperative nausea and vomiting with the application of an artificial neural network. *British Journal of Anaesthesia*, 98(1), 60–65. doi:10.1093/bja/ael282
- Penny, K. I., & Smith, G. D. (2012). The use of data-mining to identify indicators of health-related quality of life in patients with irritable bowel syndrome. *Journal of Clinical Nursing*, 21(19pt20), 2761-2771. doi:10.1111/j.1365-2702.2011.03897.x
- Petalidis, L. P., Oulas, A., Backlund, M., Wayland, M. T., Liu, L., Plant, K., ... Collins, V. P. (2008). Improved grading and survival prediction of human astrocytic brain tumours by artificial neural network analysis of gene expression microarray data. *Molecular Cancer Therapeutics*, 7(5), 1013–1024. doi:10.1158/1535-7163.MCT-07-0177
- Rajanayagam, J., Frank, E., Shepherd, R. W., & Lewindon, P. J. (2013). Artificial neural network is highly predictive of outcome in paediatric acute liver failure. *Pediatric Transplantation*, 17(6), 535-542. doi:10.1111/petr.12100
- Rezaei-Darzi, E., Farzadfar, F., Hashemi-Meshkini, A., Navidi, I., Mahmoudi, M., Varmaghani, M...Mohammad, K. (2014). Comparison of two data mining techniques in labeling diagnosis to Iranian pharmacy claim dataset: Artificial neural network (ANN) versus decision tree model. *Archives of Iranian Medicine*, 17(12), 837-843.
doi:0141712/AIM.0010
- Robinson, C. J., Swift, S., Johnson, D. D., Almeida, J. S. (2008). Prediction of pelvic organ prolapse using an artificial neural network. *American Journal of Obstetrics and Gynecology*, 199(2), 193.e1-193.e6. doi:10.1016/j.ajog.2008.04.029

- Rockwood, K., Richard, M., Leibman, C., Mucha, L., & Mitnitski, A. (2013). Staging dementia from symptom profiles on a care partner website. *Journal of Medical Internet Research, 15*(8), e145. doi:10.2196/jmir.2461
- Rodriguez-Luna, H., Vargas, H. E., & Rakela, J. (2005). Artificial neural network and tissue genotyping of hepatocellular carcinoma in liver-transplant recipients: Prediction of recurrence. *Transplantation, 79*(12), 1737-1740.
doi:10.1097/01.TP.0000161794.32007.D1
- Sachdeva, J., Kumar, V., Gupta, I., Khandelwal, N., & Ahuja, C. K. (2012). A dual neural network ensemble approach for multiclass brain tumor classification. *International Journal for Numerical Methods in Biomedical Engineering, 28*(11), 1107-1120.
doi:10.1002/cnm.2481
- Sachdeva, J., Kumar, V., Gupta, I., Khandelwal, N., & Ahuja, C. (2013). Segmentation, feature extraction, and multiclass brain tumor classification. *Journal of Digital Imaging, 26*(6), 1141-1150. doi:10.1007/s10278-013-9600-0
- Sahin, C., Ogulata, S. N., Aslan, K., Bozdemir, H., & Erol, R. (2008). A neural network-based classification model for partial epilepsy by EEG signals. *International Journal of Pattern Recognition & Artificial Intelligence, 22*(5), 973-985. doi:10.1142/s0218001408006594
- Salgueiro, M., Basogain, X., Collado, A., Torres, X., Bilbao, J., Doñate, F., & ... Azkue, J. J. (2013). Original Research Articles: An Artificial Neural Network Approach for Predicting Functional Outcome in Fibromyalgia Syndrome after Multidisciplinary Pain Program. *Pain Medicine, 14*(10), 1450-1460. doi:10.1111/pme.12185
- Salomoni, G., Grassi, M., Mosini, P., Riva, P., Cavedini, P., & Bellodi, L. (2009). Artificial neural network model for the prediction of obsessive-compulsive disorder treatment

response. *Journal of Clinical Psychopharmacology*, 29(4), 343-349.

doi:10.1097/JCP.0b013e3181aba68f

Saraoğlu, H., Temurtas, F., & Altıkat, S. (2013). Quantitative classification of HbA1C and blood glucose level for diabetes diagnosis using neural networks. *Australasian Physical & Engineering Sciences in Medicine*, 36(4), 397-403. doi:10.1007/s13246-013-0217-x

Sarbaz, Y., Towhidkhah, F., Gharibzadeh, S., & Jafari, A. (2012). Gait spectral analysis: An easy fast quantitative method for diagnosing Parkinson's disease. *Journal of Mechanics in Medicine and Biology*, 12(3), 1250041-1-1250041-13. doi:10.1142/S0219519411004691

Scheffer, C., & Cloete, T. (2012). Inertial motion capture in conjunction with an artificial neural network can differentiate the gait patterns of hemiparetic stroke patients compared with able-bodied counterparts. *Computer Methods in Biomechanics and Biomedical Engineering*, 15(3), 285-294. doi:10.1080/10255842.2010.527836

Serpen, G., Tekkedil, D., & Orra, M. (2008). A knowledge-based artificial neural network classifier for pulmonary embolism diagnosis. *Computers In Biology and Medicine*, 38(2), 204-220. doi:10.1016/j.compbiomed.2007.10.001

Sinha, R., Aggarwal, Y., & Das, B. (2007). Backpropagation artificial neural network classifier to detect changes in heart sound due to mitral valve regurgitation. *Journal of Medical Systems*, 31(3), 205-209. doi:10.1007/s10916-007-9056-1

Somfai, G. M., Tátrai, E., Laurik, L., Varga, B., ölvédy, V., Jiang, H., & ... DeBuc, D. C. (2014). Automated classifiers for early detection and diagnosis of retinopathy in diabetic eyes. *BMC Bioinformatics*, 15(1), 1-18. doi:10.1186/1471-2105-15-106

- Song, J., Lee, J., Choi, J., & Chun, S. (2013). Automatic differential diagnosis of pancreatic serous and mucinous cystadenomas based on morphological features. *Computers in Biology and Medicine*, 43(1), 1-15. doi:10.1016/j.combiomed.2012.10.009
- Song, J. H., Venkatesh, S. S., Conant, E. A., Arger, P. H., & Sehgal, C. M. (2005). Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. *Academic Radiology*, 12(4), 487-495. doi:10.1016/j.acra.2004.12.016
- Song, X., Mitnitski, A., MacKnight, C., & Rockwood, K. (2004). Assessment of individual risk of death using self-report data: an artificial neural network compared with a frailty index. *Journal of the American Geriatrics Society*, 52(7), 1180-1184. doi:10.1111/j.1532-5415.2004.52319.x
- Streba, C., Ionescu, M., Gheonea, D., Sandulescu, L., Ciurea, T., Saftoiu, A., ...Rogoveanu, I. (2012). Contrast-enhanced ultrasonography parameters in neural network diagnosis of liver tumors. *World Journal of Gastroenterology*, 18(32), 4427-4434. doi:10.3748/wjg.v18.i32.4427
- Suzuki, K. (2009). A supervised “lesion-enhancement” filter by use of a Massive-Training Artificial Neural Network (MTANN) in Computer-Aided Diagnosis (CAD). *Physics in Medicine and Biology*, 54(18), S31–S45. doi:10.1088/0031-9155/54/18/S03
- Tagluk, M., & Sezgin, N. (2010). Classification of sleep apnea through sub-band energy of abdominal effort signal using wavelets + neural networks. *Journal of Medical Systems*, 34(6), 1111-1119. doi:10.1007/s10916-009-9330-5
- Tan, M., Deklerck, R., Cornelis, J., & Jansen, B. (2013). Phased searching with NEAT in a time-scaled framework: Experiments on a computer-aided detection system for lung

- nodules. *Artificial Intelligence in Medicine*, 59(3), 157-167.
doi:10.1016/j.artmed.2013.07.002
- Tang, H., Poynton, M., Hurdle, J., Baird, B., Koford, J., & Goldfarb-Rumyantzev, A. (2011). Predicting three-year kidney graft survival in recipients with systemic lupus erythematosus. *ASAIO Journal*, 57(4), 300-309. doi:10.1097/MAT.0b013e318222db30
- Taşdelen, B., Helvacı, S., Kaleağası, H., & Özge, A. (2009). Artificial neural network analysis for prediction of headache prognosis in elderly patients. *Turkish Journal of Medical Sciences*, 39(1), 5-12. doi:10.3906/sag-0709-31
- Tejera, E., Jose Areias, M., Rodrigues, A., Ramoa, A., Manuel Nieto-Villar, J., & Rebelo, I. (2011). Artificial neural network for normal, hypertensive, and preeclamptic pregnancy classification using maternal heart rate variability indexes. *Journal of Maternal-Fetal & Neonatal Medicine*, 24(9), 1147-1151, doi: 10.3109/14767058.2010.545916
- Tokuda, O., Harada, Y., Ohishi, Y., Matsunaga, N., & Edenbrandt, L. (2014). Investigation of computer-aided diagnosis system for bone scans: A retrospective analysis in 406 patients. *Annals of Nuclear Medicine*, 28(4), 329-339. doi:10.1007/s12149-014-0819-8
- Tseng, W., Hung, L., Shieh, J., Abbod, M. F., & Lin, J. (2013). Hip fracture risk assessment: Artificial neural network outperforms conditional logistic regression in an age- and sex-matched case control study. *BMC Musculoskeletal Disorders*, 14(207).
doi:10.1186/1471-2474-14-207
- Uğuz, H. (2012). A Biomedical System Based on Artificial Neural Network and Principal Component Analysis for Diagnosis of the Heart Valve Diseases. *Journal of Medical Systems*, 36(1), 61-72. doi:10.1007/s10916-010-9446-7

- Waller, T., Nowak, R., Tkacz, M., Zapart, D., & Mazurek, U. (2013). Familial or Sporadic Idiopathic Scoliosis - classification based on artificial neural network and GAPDH and ACTB transcription profile. *Biomedical Engineering Online*, 12(1), 1-14.
doi:10.1186/1475-925X-12-1
- Wang, C., Li, L., Wang, L., Ping, Z., Flory, M. T., Wang, G., . . . Li, W. (2013). Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach. *Diabetes Research and Clinical Practice*, 100(1), 111-118.
doi:10.1016/j.diabres.2013.01.023
- Wang, X., Lederman, D., Tan, J., Wang, X. H., & Zheng, B. (2011). Computerized prediction of risk for developing breast cancer based on bilateral mammographic breast tissue asymmetry. *Medical Engineering & Physics*, 33(8), 934–942.
doi:10.1016/j.medengphy.2011.03.001
- Witt, D. R., Chen, H., Mielens, J. D., McAvoy, K. E., Zhang, F., Hoffman, M. R., & Jiang, J. J. (2014). Detection of chronic laryngitis due to laryngopharyngeal reflux using color and texture analysis of laryngoscopic images. *Journal of Voice*, 28(1), 98-105.
doi:10.1016/j.jvoice.2013.08.015
- Yalcin, N., Tezel, G., & Karakuzu, C. (2015). Epilepsy diagnosis using artificial neural network learned by PSO. *Turkish Journal of Electrical Engineering and Computer Sciences*, 23(2), 421-432. doi:10.3906/elk-1212-151
- Yang, S., Nam, Y., Kim, M., Kim, E. Y., Park, J., & Kim, D. (2013). Computer-aided detection of metastatic brain tumors using magnetic resonance black-blood imaging. *Investigative Radiology*, 48(2), 113-119. doi:10.1097/RLI.0b013e318277f078

Yu, Z., Lu, H., Si, H., Liu, S., Li, X., Gao, C., & ... Yao, X. (2015). A Highly Efficient Gene Expression Programming (GEP) Model for Auxiliary Diagnosis of Small Cell Lung Cancer. *Plos One*, *10*(5), doi:10.1371/journal.pone.0125517

Zhang, X., Kanematsu, M., Fujita, H., Zhou, X., Hara, T., Yokoyama, R., & Hoshi, H. (2009). Application of an artificial neural network to the computer-aided differentiation of focal liver disease in MR imaging. *Radiological Physics and Technology*, *2*(2), 175-182. doi:10.1007/s12194-009-0062-5

Appendix H

Weight of Evidence Analysis Matrix

WEIGHT OF EVIDENCE ANALYSIS		WoE-A			WoE-B		WoE-C			WoE-D	Accept?	Count: 127, Accepted: 74 (58.27%)	
Study ID	Author/Date	# Cases	Validation	D/P Appl	SCORE	ANN Study	SCORE	ANN Detail	ANN Perf	SCORE	TOTAL	Yes/No	Comments
3	Ahmed et al., 2013	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
5	Yu et al., 2015	Yes	No	Yes	2	Some	1	Yes	Yes	3	6	No	Gene Express Programming, not ANN focused
6	Ozbay, 2009										1	No	Complex Discrete Wavelet
7	Avci et al., 2014										1	No	This study involves non-patient sampling - does not meet criteria
8	Erol et al., 2008	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
9	Salgueiro et al., 2013	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
11	Niwas et al., 2012										1	No	Complex Discrete Wavelet
12	Mei-Ling et al., 2006	Yes	Some	Yes	2	Yes	3	Yes	Some	2	7	Yes	
14	Chen & Chou, 2014	Some	Some	Yes	2	Yes	3	Yes	Yes	3	8	Yes	
16	Tejera et al., 2011	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
17	Chiu et al., 2009	Some	Some	Yes	2	Yes	3	Some	Yes	2	7	Yes	
19	Aggarwal et al., 2007	Yes	Some	Yes	2	Yes	3	Yes	Yes	3	8	Yes	
20	Mert et al., 2015	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
21	Kocycigit et al., 2008	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
23	Tagluk & Sezgin, 2010	No	Some	Yes	1	Yes	3	Yes	Yes	3	7	No	Too few cases (21)
31	Olsson et al., 2006	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
33	Delavarian et al., 2012	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
35	Mariani et al., 2012	No	Yes	No	1	Yes	3	Yes	Yes	3	7	No	Non-Diagnostic (sleep cycle)
38	Yalcin et al., 2015	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
43	Diao et al., 2011	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
44	Liu et al., 2009	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
47	La Torre et al., 2010	Yes	Yes	Yes	3	Some	2	Yes	Yes	3	8	Yes	
49	Fraschini, 2011	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
52	Kara & Guven, 2007	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
53	Lweesy et al., 2011	Some	Yes	Yes	2	Yes	3	Some	Yes	2	7	Yes	
55	Kshirsagar et al., 2006	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
56	Kumar & Sahoo, 2013	Yes	Yes	Yes	3	Some	1	No	Yes	2	6	No	Focus on Rules-based support method, not ANN
62	Aydin et al., 2011	No	Yes	Yes	1	Some	2	Some	Yes	2	5	No	The study split the sample of 30 into 3 independent groups, thus effectively making them samples of 10
63	Penny & Smith, 2012	Yes	Yes	No	1	Some	2	Some	Yes	2	5	No	Focus on quality of life with IBS, not diag/prog
64	Huang et al., 2012										1	No	Article could not be retrieved from databases
65	Rezaei-Darzi et al, 2014	No	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	Study of population health regarding pharmaceutical vs diagnosis
66	Farhadian et al, 2015	Yes	Yes	No	1	Yes	3	Some	Yes	2	6	No	Focus on disease prediction, not diag/prog
70	Lux et al, 2013	Yes	Some	Yes	2	Some	2	Some	Yes	2	6	No	Insufficient ANN detail and validation info
77	Ibrahim et al, 2010	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
83	Andersson et al, 2013	Yes	No	Yes	1	Yes	3	Some	Yes	2	6	No	Insufficient ANN detail and validation info
93	Michalski et al, 2011										1	No	Replication of article #94, intentionally bypassed
94	Michalski et al, 2011	Yes	No	Yes	1	No	1	Some	No	2	4	No	The ANN was used as a classifier, not outcome
95	Castillo et al, 2006	No	Some	Yes	1	Yes	3	Some	Yes	2	6	No	Insufficient cases, ANN detail, and validation info
99	Cheng et al, 2013	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
100	Tasdelen et al, 2009	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	

WEIGHT OF EVIDENCE ANALYSIS		WoE-A			WoE-B		WoE-C			WoE-D	Accept?	Count: 127, Accepted: 74 (58.27%)	
Study ID	Author/Date	# Case	Validation	D/P Appl	SCORE	ANN Study	SCORE	ANN Detail	ANN Perf	SCORE	TOTAL	Yes/No	Comments
103	Kaya, 2014	Yes	Some	Yes	2	Yes	3	Yes	Yes	3	8	Yes	
108	Tseng et al, 2013	Yes	Yes	No	1	Yes	3	Yes	Yes	3	7	No	Risk-factor assessment, not Diag/Prog focused
112	Somjai et al., 2014	Yes	Some	Yes	2	Yes	3	Yes	Some	2	7	Yes	Validation was done post-study, not with training
121	Rajanayagam et al, 2013	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
124	Cho et al, 2013	No	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	Low number of sample cases available (25)
133	Gargouri et al, 2012	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
134	Waller et al, 2013	No	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	Low number of sample cases available (29)
141	Mielens et al, 2012	No	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	Low number of sample cases available (25)
161	Sarbaz et al, 2012	Yes	Yes	Yes	2	Yes	3	Yes	Yes	2	7	Yes	
164	Sahin et al, 2008	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
165	Bollschweiler et al, 2004	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
166	Hallner&Hasenbring, 2004	Yes	Some	Yes	2	Yes	3	Yes	Yes	3	8	Yes	
169	Song et al, 2004	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
171	Joo et al, 2004	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
172	Fernandez et al, 2005	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
173	Su et al, 2005										1	No	Article could not be retrieved from databases
174	Bhatikar et al, 2005	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
175	Song et al, 2005	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
177	Rodriguez-Luna et al, 2005	No	No	Yes	1	Yes	2	No	Yes	2	5	No	Insufficient cases and no validation described
178	Kewk et al, 2005	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
179	Alkan et al, 2005	No	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	Insufficient cases (11 patients, even tho many samples)
181	Nayak et al, 2006	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
182	Pellegrini et al, 2005	Yes	Yes	Yes	3	Yes	2	Some	Some	2	7	Yes	
184	Mofidi et al, 2006	Yes	Yes	Yes	3	Yes	3	Yes	Yes	2	8	Yes	
185	Behrman et al, 2006	Yes	Yes	Yes	3	Yes	3	Yes	Some	2	8	Yes	
186	Acton&Newberg, 2006	Yes	Some	Yes	1	Yes	3	Some	Yes	2	6	No	All data was used for both training and test - validity suspect
187	Limonaldi et al, 2006	Yes	Yes	Yes	3	Yes	3	Yes	Yes	2	8	Yes	
188	Amodio et al, 2006				1		1			1	3	No	Replicate article (French) to #182, intentionally bypassed
189	Chun et al, 2007	Yes	No	Yes	2	Yes	3	No	Yes	1	6	No	ANN selected/used was not detailed in this study
190	Peng et al, 2006	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
191	Linder et al, 2008	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
197	Carrara et al, 2007	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
200	Sinha et al, 2007	No	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	Only 20 subjects involved in the study
201	Norman et al, 2007	No	Yes	Yes	1	Yes	3	Some	Yes	2	6	No	Only 18 subjects involved in the study (although multiple samples were taken from each subject)
203	Serpen et al, 2008	Yes	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	The study cases were simulations, not actual study data - hard to assure generalization from findings
204	Lin, 2008	Yes	Yes	Yes	2	Yes	3	Some	Yes	2	7	Yes	
205	Kocyigit et al, 2008				1		1			1	3	No	Replicate article to #21, intentionally bypassed
206	Petalidis et al, 2008	Yes	Some	Yes	2	Yes	3	Some	Unsure	1	6	No	The performance detail for training set was given but unclear on performance reported for testing
207	Robinson et al, 2008	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
212	Grewal et al, 2008										1	No	Article could not be retrieved from databases
215	McLaren et al, 2009	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
216	Salomoni et al, 2009	Yes	Yes	Yes	2	Yes	3	Yes	Yes	3	8	Yes	Validation of network was limited

WEIGHT OF EVIDENCE ANALYSIS		WoE-A				WoE-B		WoE-C			WoE-D	Accept?	Count: 127, Accepted: 74 (58.27%)
Study ID	Author/Date	# Cases	Validation	D/P Appl	SCORE	ANN Study	SCORE	ANN Detail	ANN Perf	SCORE	TOTAL	Yes/No	Comments
218	Abouzari et al, 2009	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
219	Suzuki, 2009	Yes	Yes	Yes	3	Yes	3	Yes	Yes	1	7	No	ANN application used for pixel enhancement, not diagnostic determination
227	Kocer & Canal, 2009	No	Yes	Yes	1	Yes	3	Some	Yes	2	6	No	Difficult to ascertain number of cases used, and ANN applied across multiple algorithms needed further detail
230	Uguz, 2012	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
231	Isik & Arslan, 2011	Yes	Some	Yes	2	Partial	2	Yes	Some	2	6	No	Applies ANN within diagnostic process and not for direct diagnostic decision making
233	Hui et al, 2011	Yes	Yes	Yes	3	Yes	2	Yes	Yes	3	8	Yes	
234	Zhang et al, 2009	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
241	Park et al, 2011	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
242	Linder et al, 2011										1	No	Article could not be retrieved from databases
245	Dietzel et al, 2012	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
247	Scheffer & Cloete, 2012	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
248	Wang et al, 2011	Yes	Yes	Yes	3	Yes	3	Yes	Yes	2	8	Yes	
250	Tang et al, 2011	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
254	Saftoiu et al, 2012										1	No	Article could not be retrieved from databases
255	Dey et al, 2011	Yes	Yes	Yes	2	Yes	3	Some	Yes	2	7	Yes	
264	Biglarian et al, 2012	Yes	Yes	Yes	3	Yes	3	Some	Some	2	8	Yes	AUROC in Abstract not mentioned in actual study
268	Streba et al, 2012	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
272	Hoffman et al, 2013	No	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	Only 25 subjects included in the study
273	Sachdeva et al, 2012	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
275	Halford et al, 2013	Yes	Yes	Yes	3	Yes	2	Some	Some	1	6	No	Performance analysis was across interrater scorers - too complex for this analysis
276	Song et al, 2013	No	Yes	Yes	1	Yes	3	Some	Yes	2	6	No	Only 7 subjects included in the study (11 samples)
277	Yang et al, 2013	No	Yes	Yes	1	Partial	2	Some	Some	2	5	No	Only 26 subjects included in the study (128 samples)
286	Ventouras et al, 2012										1	No	Article could not be retrieved from databases
288	Geng et al, 2013	Yes	Yes	Yes	3	Partial	2	Yes	Yes	3	8	Yes	Examined new manometry technique applying an ANN
289	Wang et al, 2013	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
292	Agular-Pulido et al, 2013										1	No	Article could not be retrieved from databases
296	Sachdeva et al, 2013	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
299	Hoffman et al, 2013a	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	Different study than #272 above (30 subjects included)
305	Rockwood et al, 2013	Yes	Yes	Yes	3	Yes	3	No	Yes	1	7	No	There was no detail presented on the ANN model's structure or design - resulted in low confidence
307	Hossen, 2013	Yes	Yes	Yes	2	Yes	3	Some	Unsure	1	6	No	Much of the analysis was difficult to assess due to poor translation of the study - not including in final review
309	Saraoglu et al, 2013	Yes	Yes	Yes	2	Yes	3	Some	Unsure	1	6	No	Non-standard measure used to assess ANN performance (mean relative absolute error) which could not be used here
311	Tan et al, 2013	Yes	Yes	Yes	3	Partial	1	Yes	yes	2	6	No	This study compares various ANN algorithm implementations which does not serve our analysis
316	Casti et al, 2013										1	No	Article could not be retrieved from databases
319	Kuruville & Gunavathi, 2014	Some	Yes	Yes	2	Yes	3	Yes	Yes	2	7	Yes	Detail given for algorithm selection, but minimum on ANN application used (commercial, custom, other)

WEIGHT OF EVIDENCE ANALYSIS		WoE-A				WoE-B		WoE-C			WoE-D	Accept?	Count: 127, Accepted: 74 (58.27%)
Study ID	Author/Date	# Cases	Validation	D/P Appl	SCORE	ANN Study	SCORE	ANN Detail	ANN Perf	SCORE	TOTAL	Yes/No	Comments
322	Matulewicz et al, 2014	No	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	Only 18 subjects involved in the study (although about 5500 MRI "voxels" [graphical samples] were examined)
323	Bevevino et al, 2014	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
326	Witt et al, 2014	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
329	Lai et al, 2014	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	
332	Pachori & Patidar, 2014	No	Yes	Yes	1	Yes	3	Yes	Yes	3	7	No	Only 5 subjects involved in the study (although each had 100 channel EEGs examined)
337	He et al, 2014	Yes	Yes	Yes	3	Yes	4	Some	Yes	2	9	Yes	
343	Tokuda et al, 2014	Yes	Some	Yes	1	Yes	3	Some	Yes	2	6	No	Minimal description of ANN design and detail
344	Jovanovic et al, 2014	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	ERCP is an endoscopy procedure for the upper GI tract
351	Mohamed et al, 2014	Some	Some	Yes	1	Yes	3	Some	Yes	2	6	No	Insufficient cases and validation description
353	Gheonea et al, 2014	Yes	Yes	Yes	3	Yes	3	Yes	Yes	3	9	Yes	
361	Li et al, 2014	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes	

Appendix I

Research Question Analysis Matrix

Study ID#	--- RQ#1 ---		OUTCOME	--- RQ#2 ---						APPLICATION	Comments
	PERFORMANCE	PERFORMANCE		METHODOLOGY			APPLICATION				
	Measure	Value	Classification Successful?	Study Type	Primary Tool	Hybridized with...	Algorithm Specified	Comparative Analysis	ANN Best	Disease/Issue Application	
3	Prediction %	98.10%	Yes	Diagnostic	MatLab	MPANN	Genetic	No	n/a	Breast Cancer	Strong manipulation of ANN parameters for optimization
8	Prediction %	91.60%	Yes	Diagnostic	MLPBPNN	n/a	Leven-Marq	Yes	No	Thyroid Disease	Alternate methodology had somewhat greater success
9	Prediction %	88.38%	Yes	Prognostic	MLPBPNN	n/a	n/a	Yes	Yes	Fibromyalgia Syndrome	Compared favorably to Logistic Regression across 3 scenarios
12	AUROC	94.90%	Yes	Diagnostic	MLPBPNN	n/a	n/a	Yes	Yes	Glaucoma	Compared favorably to Logistic Regression
14	AUROC	94.16%	Yes	Diagnostic	MLPBPNN	n/a	CG	Yes	Yes	Chest Auscultation	Performance compared similarly to Learning Vector Quantization (LVQ)
16	AUROC	97.27%	Yes	Diagnostic	SPSS	LZ	n/a	No	n/a	Gestational Heart Rate	Specificity not strong in Preclampsia patients
17	AUROC	88.00%	Yes	Prognostic	StatSoft	n/a	n/a	No	n/a	Skeletal Metastasis of PC	
19	Prediction %	98.03%	Yes	Diagnostic	C++	n/a	n/a	No	n/a	Exertional Heat Illness	Home-grown ANN developed using C++ programming
20	Prediction %	99.12%	Yes	Diagnostic	MLPBPNN	ICA	n/a	No	n/a	Breast Cancer	Augmentation by ICA was evaluated for overall performance improvement beyond machine learning alone (multiple)
21	Sens-Spec	98.00%/90.50%	Yes	Diagnostic	MLPBPNN	ICA	n/a	No	n/a	Epilepsy	
31	AUROC	98.00%	Yes	Diagnostic	MLPBPNN	n/a	Langevin	No	n/a	Transmural Ischemia	
33	Prediction %	95.50%	Yes	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	ADHD	MLPNN shown slightly less accurate than RBFNN
38	Prediction %	99.67%	Yes	Diagnostic	MLPBPNN	PSO	n/a	No	n/a	Epilepsy	
43	Sens-Spec	100.00%/80.80%	Yes	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Breast Cancer	
44	AUROC	99.00%	Yes	Prognostic	MatLab	n/a	Bayesian	No	n/a	Pneumoconiosis	
47	Prediction %	64.00%	Partial	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Melanoma	ANN outperformed by SVM in this study, but it beat k-NN
49	AUROC	91.00%	Yes	Diagnostic	MLPBPNN	Wavelet	n/a	No	n/a	Breast Cancer	Focus on ROI detection rather than cancer/non-cancer
52	Prediction %	96.77%	Yes	Diagnostic	MatLab	n/a	Leven-Marq	No	n/a	Optic Nerve Disease	
53	Prediction %	92.30%	Yes	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Obstructive Sleep Apnea	Diagnostic input from ECG wave forms
55	AUROC	72.50%	Partial	Diagnostic	neURON++	n/a	n/a	No	n/a	Hypogonadism	ANN outperformed linear analyses (LR, LDFA, QDFA)
77	Prediction %	96.86%	Yes	Prognostic	MLPBPNN	n/a	n/a	No	n/a	Dengue Fever Risk	
99	AUROC	98.10%	Yes	Diagnostic	MLPBPNN	PSO	Genetic	No	n/a	Basal Cell Carcinoma	Final analysis was across ANN ensemble
100	AUROC	83.35%	Yes	Prognostic	StatSoft	n/a	n/a	No	n/a	Headaches	AUROC is averaged across three periodic study components
103	Prediction %	96.79%	Yes	Diagnostic	MLPBPNN	ELM	n/a	No	n/a	Thyroid Disease	
112	Prediction %	90.00%	Yes	Diagnostic	MatLab	n/a	Markov Chain	No	n/a	Diabetic Retinopathy	Performance limited to two of four studies
121	AUROC	96.00%	Yes	Prognostic	WEKA	n/a	n/a	No	n/a	Acute Liver Failure	
133	AUROC	95.00%	Yes	Diagnostic	MLPBPNN	n/a	n/a	Yes	Yes	Breast Cancer	

Count: 74

Coded from full set: 58.27%

Study ID#	--- RQ#1 ---			--- RQ#2 ---						Count: 74	
	PERFORMANCE		OUTCOME	METHODOLOGY						APPLICATION	Comments
	Performance Measure	Performance Value	Classification Successful?	Study Type	Primary Tool	Hybridized with...	Algorithm Specified	Comparative Analysis	ANN Best	Disease/Issue Application	
161	Prediction %	92.86%	Fully	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Parkinson's Disease	
164	Prediction %	96.00%	Fully	Diagnostic	MLPBPNN	n/a	Leven-Marq	No	n/a	Epilepsy	
165	Prediction %	93.00%	Fully	Diagnostic	MLPBPNN	n/a	n/a	Yes	Yes	Metastatic Cancer	
166	Prediction %	83.10%	Fully	Prognostic	MLPBPNN	n/a	n/a	No	n/a	Psychosocial Risk	
169	AUROC	86.00%	Fully	Prognostic	MLPBPNN	n/a	n/a	Yes	Yes	Risk of Death	
171	AUROC	95.00%	Fully	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Breast Cancer	
172	Sens-Spec	75.00./92.50%	Partial	Diagnostic	MatLab	n/a	n/a	Yes	Yes	Hemodialysis	
174	Sens-Spec	88.00%/83.00%	Fully	Diagnostic	CU-ANN	n/a	n/a	No	n/a	Heart Murmur	
175	AUROC	85.60%	Fully	Diagnostic	MLPBPNN	n/a	n/a	Yes	Equal	Breast Cancer	Not a significant outcome difference, although ANN slightly better than LR
178	Sens-Spec	99.20%/99.40%	Fully	Diagnostic	MatLab	n/a	n/a	No	n/a	Colorectal Cancer	
181	Prediction %	98.30%	Fully	Diagnostic	MatLab	n/a	n/a	Yes	Yes	Oral Cancer	Only a small outcome difference, although ANN slightly better than PCA
182	Spearman	84.00%	Fully	Diagnostic	ANNES	Expert	n/a	No	n/a	Hepatic Encephalopathy	Performance of ANN acceptable except for one testing class (no agreement with expert)
184	Prediction %	89.75%	Fully	Prognostic	Neurosol	n/a	n/a	No	n/a	Esophageal Cancer	
185	Prediction %	69.00%	Partial	Diagnostic	MLPBPNN	n/a	Adapt Prop	Yes	Equal	Neuropathic Pain	ANN performed slightly better than LR, but neither was definitive
187	Prediction %	95.00%	Partial	Diagnostic	Neurosol	n/a	n/a	No	n/a	Facial Pain	ANN performed extremely well for the major category of cases, but for others it was good-to-poor
190	Prediction %	83.30%	Fully	Prognostic	StatSoft	n/a	n/a	No	n/a	Post-op Nausea	
191	Prediction %	80.00%	Partial	Diagnostic	MLPBPNN	n/a	Adapt Prop	No	n/a	Disphonia	ANN performed well for one class but mediocre for several others
197	Sens-Spec	88.00%/80.00%	Fully	Diagnostic	StatSoft	n/a	CG	No	n/a	Melanoma	
204	Prediction %	83.00%	Fully	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Scoliosis	
207	Sens-Spec	90.00%/83.00%	Fully	Prognostic	MatLab	n/a	n/a	No	n/a	Pelvic Organ Prolapse	
215	AUROC	87.00%	Fully	Diagnostic	MLPBPNN	n/a	n/a	Yes	Equal	Breast Cancer	
216	AUROC	94.50%	Fully	Prognostic	MLPBPNN	n/a	n/a	Yes	Yes	Obsessive-Compulsive Disorder	
218	AUROC	76.70%	Partial	Prognostic	MatLab	n/a	n/a	Yes	Yes	Chronic Subdural Hematoma	
230	Prediction %	95.00%	Fully	Diagnostic	MatLab	PCA	n/a	No	n/a	Heart Valve Disease	
233	AUROC	73.70%	Partial	Prognostic	EasyNN	n/a	n/a	Yes	Equal	Febrile Neutropenia	
234	Prediction %	93.30%	Fully	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Focal Liver Disease	
241	AUROC	88.40%	Fully	Diagnostic	MLPBPNN	GA	n/a	No	n/a	Interstitial Lung Disease	
245	AUROC	88.80%	Fully	Diagnostic	Math Works	n/a	n/a	No	n/a	Breast Cancer	
247	Prediction %	99.40%	Fully	Diagnostic	MatLab	n/a	Leven-Marq	No	n/a	Hemiparetic Stroke	
248	AUROC	78.10%	Partial	Prognostic	MLPBPNN	GA	n/a	No	n/a	Breast Cancer	
250	AUROC	71.00%	Partial	Prognostic	MLPBPNN	n/a	n/a	Yes	Equal	Kidney Graft	ANN performance was exceeded by LR but not significantly
255	Prediction %	93.75%	Fully	Diagnostic	Neurointell	n/a	n/a	No	n/a	Breast Cancer	Prediction % calculated from raw data
264	Prediction %	78.45%	Partial	Prognostic	MLPBPNN	n/a	n/a	Yes	Yes	Metastatic Cancer	Prediction % calculated from raw data
268	Prediction %	87.12%	Fully	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Liver Cancer	A bit weak on ANN detail

Study ID#	--- RQ#1 ---			--- RQ#2 ---						Count: 74	
	PERFORMANCE		OUTCOME	METHODOLOGY						APPLICATION	Coded from full set: 58.27%
	Performance Measure	Performance Value	Classification Successful?	Study Type	Primary Tool	Hybridized with...	Algorithm Specified	Comparative Analysis	ANN Best	Disease/Issue Application	Comments
273	Prediction %	90.40%	Fully	Diagnostic	MatLab	PCA	n/a	No	n/a	Brain Cancer	
288	AUROC	90.90%	Fully	Diagnostic	MatLab	n/a	n/a	No	n/a	Swallowing disorder	AUROC is averaged across two ANN versions employed (both successful)
289	AUROC	89.10%	Fully	Prognostic	MatLab	n/a	n/a	Yes	Yes	Diabetes Mellitus	
296	Prediction %	85.23%	Fully	Diagnostic	MatLab	PCA	n/a	No	n/a	Brain Cancer	
299	AUROC	95.64%	Fully	Diagnostic	MatLab	n/a	n/a	No	n/a	Swallowing disorder	AUROC is averaged across 12 ANN versions, all successful
319	Prediction %	91.11%	Fully	Diagnostic	MLPBPNN	n/a	GD	No	n/a	Lung Cancer	
323	AUROC	80.00%	Fully	Prognostic	MLPBPNN	n/a	n/a	Yes	Yes	Limb Fracture - Open	
326	AUROC	88.70%	Fully	Diagnostic	MLPBPNN	n/a	Leven-Marq	No	n/a	Laryngopharyngeal Reflux	
329	Prediction %	99.00%	Fully	Diagnostic	MLPBPNN	n/a	GD	No	n/a	Autoimmune Hemolytic Anemia	
337	Prediction %	73.05%	Partial	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Lung Cancer	2 ANNs were used, performance was averaged (though both were within 1.5% of each other)
344	AUROC	88.40%	Fully	Prognostic	SPSS	n/a	n/a	Yes	Yes	Endoscopic Therapy	ERCP (endoscopic therapy) has risks, and this attempts to mitigate that risk given disease presentation.
353	Sens-Spec	95.09%/92.19%	Fully	Diagnostic	MLPBPNN	n/a	n/a	No	n/a	Liver Cancer	Sens-Spec value was averaged between two individual ANNs employed in the research (both met the "Classification Successful?" criteria)
361	Prediction %	88.00%	Fully	Diagnostic	WEKA	n/a	n/a	Yes	No	Metastatic Cancer	Naïve Bayesian classifier performed better than ANN and k-NN classifiers, primarily on sensitivity

Appendix J

Coding Examples

Weight of Evidence Analysis - Andersson, Heijl, Bizios, and Bengtsson (2013)

WEIGHT OF EVIDENCE ANALYSIS		WoE-A				WoE-B		WoE-C			WoE-D	Accept?	Comments
Study ID	Author/Date	# Cases	Validation	D/P Applic	SCORE	ANN Study	SCORE	ANN Detail	ANN Perf	SCORE	TOTAL	Yes/No	
83	Andersson et al, 2013	Yes	No	Yes	1	Yes	3	Some	Yes	2	6	No	

Acta Ophthalmologica
ACTA OPHTHALMOLOGICA 2013

Comparison of clinicians and an artificial neural network regarding accuracy and certainty in performance of visual field assessment for the diagnosis of glaucoma

Sabina Andersson, Anders Heijl, Dimitrios Bizios and Boel Bengtsson

Department of Clinical Sciences, Ophthalmology, Lund University, Skåne University Hospital, Malmö, Sweden

ABSTRACT.
Purpose: To compare clinicians and a trained artificial neural network (ANN) regarding accuracy and certainty of assessment of visual fields for the diagnosis of glaucoma.
Methods: Thirty physicians with different levels of knowledge and experience in glaucoma management assessed 30-2 SITA Standard visual field printouts that included full Statpac information from 99 patients with glaucomatous optic neuropathy and 66 healthy subjects. Glaucomatous eyes with perimetric mean deviation values worse than -10 dB were not eligible. The fields were graded on a scale of 1-10, where 1 indicated healthy with absolute certainty and 10 signified glaucoma; 5.5 was the cut-off between healthy and glaucoma. The same fields were classified by a previously trained ANN. The ANN output was transformed into a linear scale that matched the scale used in the subjective assessments. Classification certainty was assessed using a classification error score.
Results: Among the physicians, sensitivity ranged from 61% to 96% (mean 83%) and specificity from 59% to 100% (mean 90%). Our ANN achieved 93% sensitivity and 91% specificity, and it was significantly more sensitive than the physicians ($p < 0.001$) at a similar level of specificity. The ANN classification error score was equivalent to the top third scores of all physicians, and the ANN never indicated a high degree of certainty for any of its misclassified visual field tests.
Conclusion: Our results indicate that a trained ANN performs at least as well as physicians in assessments of visual fields for the diagnosis of glaucoma.

Key words: artificial neural network - diagnosis - glaucoma - interpretation - subjective assessment - visual field

Acta Ophthalmol. 2013; 91: 413-417
© 2012 The Authors
Acta Ophthalmologica © 2012 Acta Ophthalmologica Scandinavica Foundation
doi: 10.1111/j.1755-3768.2012.02435.x

413

Coding Example

deviation (MD) and the Glaucoma Hemifield Test (GHT) (Katz et al. 1991; Asman & Heijl 1992).

New techniques such as machine-learning classifiers and, in particular, artificial neural networks (ANNs) have been suggested for interpretation of visual field test results (Goldbaum et al. 1994, 2005; Henson et al. 1996; Chan et al. 2002; Sample et al. 2002, 2004; Tucker et al. 2005). In some cases, the ANNs in the cited studies were trained and tested with input based on threshold sensitivity values and patient age, and their performance was compared with that of glaucoma experts who had only very limited or no access to Statpac. In a previous study (Bengtsson et al. 2005), we investigated the effects of different types of perimetric test result variables on the classification performance of an ANN. We found that probability scores based on age and general height-corrected probabilities (i.e. pattern deviation probabilities) were superior to any other type of perimetric input data. This finding was confirmed in a subsequent study that used the same trained ANN and included the visual fields of an entirely independent sample of healthy and glaucomatous subjects (Bizios et al. 2007).

The output function of our ANN allows not only classification of a test but also estimation of classification certainty. Knowledge about certainty of classifications is extremely valuable for diagnosis of glaucoma. In this study, clinicians with full access to Statpac single-field analysis were compared with a fully trained ANN regarding the accuracy and certainty of classification of visual fields for the diagnosis of glaucoma.

Material and methods

Ophthalmology residents and ophthalmologists with varying experience of glaucoma who were employed at Skåne University Hospital in Lund and Malmö, Sweden, were asked to assess visual field printouts from healthy and glaucomatous individuals. All visual fields were obtained using the 30-2 SITA Standard program of the Humphrey Field Analyzer II (model 750; Carl Zeiss Meditec, Inc., Dublin, CA, USA). Field tests with fixation losses exceeding 20% or false-

positive responses >5% were excluded. Collection of the visual fields was conducted according to the tenets of the Declaration of Helsinki and was approved by the Ethics Committee of Lund University.

Patients with glaucoma

All visual fields of the patients with glaucoma were selected from the database for one of the Humphrey perimeter used at the Department of Ophthalmology, Malmö University Hospital. The selection procedure has previously been described in detail (Bizios et al. 2007). A pseudo-random selection was performed: one of every five individuals was selected, starting alphabetically and choosing in alternating order either the right or the left eye. Inclusion criteria for the patients were as follows: a diagnosis of primary open-angle glaucoma, exfoliation or pigmentary glaucoma; age 50 years or older; presence of glaucomatous optic neuropathy. Glaucomatous neuropathy was defined as optic nerve damage with structural changes in the optic disk, either detected by a glaucoma expert evaluating optic nerve head photographs or described comprehensively in patient records. Eyes with MD values worse than -10 dB were ineligible, as were patients with retinal changes, neurological or endocrinological disorders, or other conditions that could affect the visual field. Patients who had diabetes mellitus but no retinopathy and glaucoma patients with concomitant cataract were included. Visual field tests with false-positive rates > 15% or fixation losses exceeding 20% and no visible blind spot were not included. High false-negative rates were not used as criterion for exclusion, as FN rates have been shown to correlate with the amount of glaucomatous visual field loss (Katz & Sommer 1988; Bengtsson & Heijl 2000).

Healthy subjects

A random sample of visual fields of healthy subjects was retrieved from a large existing database originally created to establish normal limits for the SITA strategies (Bengtsson & Heijl 1999). For the purpose of the present study, only subjects with 50 years of age or older were included. One visual field per subject was randomly selected for inclusion.

Subjective assessment

Participating physicians assessed paper printouts of visual fields in random order. The fields were scored using a scale of 1–10 for each field test, where 1 indicated healthy with absolute certainty and 10 absolute certainty of glaucoma, with the cut-off between the healthy and glaucoma groups set at 5.5 (Fig. 1). The printout format was the single-field analysis containing full Statpac information including probability maps and Glaucoma Hemifield Test classification. The graders were masked to all other patient-related information, including subject identity, and they performed the assessments independently. Participants were also asked to indicate which of the following categories they felt applied to themselves: glaucoma specialist, general ophthalmologist, other subspecialist or resident. No statistical comparison was performed among the different subgroups of clinicians.

Assessment by an ANN

All visual fields were analysed by a fully trained ANN with the aim of identifying glaucomatous visual field loss (Bizios et al. 2007). The ANN we employed was a fully connected feed-forward multilayer perceptron that used probability scores from pattern deviation probability maps as input data (Bengtsson et al. 2005; Bizios et al. 2007). No visual fields previ-

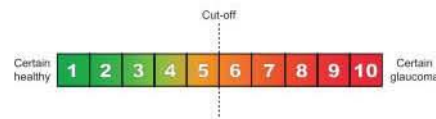
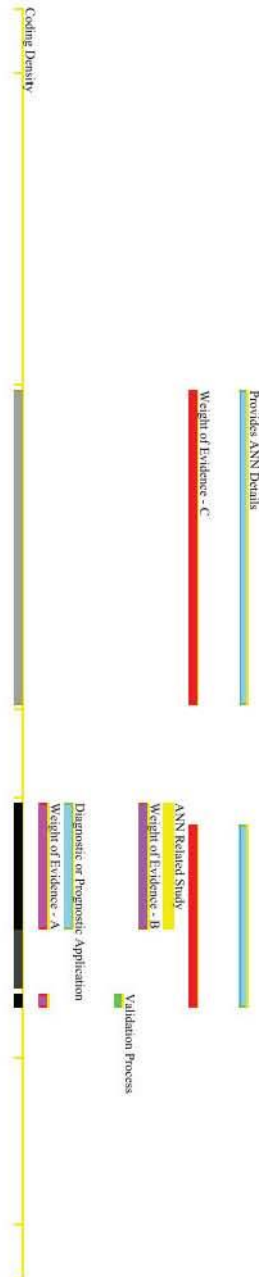


Fig. 1. The visual field assessment scale. A score of 1 indicates absolutely certain assessment as a healthy visual field, and 10 signifies certain glaucoma. The cut-off between healthy and glaucoma is at 5.5.



ously used to train the ANN were included in the analysis (Bengtsson et al. 2005).

The output of an ANN is a continuous logistic function ranging from 0 to 1, with a cut-off for the best separation between healthy and glaucoma. Values that are close to the end points 0 and 1 indicate larger certainty that a test is normal or abnormal, respectively. Values close to the cut-off still classify test results as either normal or abnormal, but with larger uncertainty. To enable comparison with the certainty in the scoring performed by the clinicians in our study, the network output was mapped to a real-valued line in MATLAB version 4.0 (The Math-Works Inc., Natick, MA, USA) and then divided into 10 equal intervals.

Analyses

Sensitivity and specificity values were calculated for the ANN, for each clinician and for the different subgroups of clinicians. Differences in sensitivity and specificity between the ANN and the clinicians were analysed using McNemar's test. The statistics were calculated by spss (version 16.0.0 for Mac; SPSS Inc., Chicago, IL, USA).

Certainty in subjective assessment was estimated using the scores obtained on the 10-grade scale. First, we counted the number of steps from the end points to the actual scores; for the visual field of a healthy eye, this was performed from the end point of healthy with absolute certainty (1); for the visual field of a glaucomatous eye, it was performed from the corresponding end point of glaucoma (10). Thereafter, for each grader, a classification error score was calculated as the mean number of steps for all fields. A classification error score for the ANN was calculated in the same way. Thus, for both the graders and the ANN, a low classification error score indicated more correct assessments with better certainty, and a high score implied the opposite. The classification error scores of the physicians and the ANN were compared using a paired *t*-test.

Results

The visual fields of 99 patients with glaucoma aged 55–95 years (mean 75.8 years) and 66 healthy individuals aged 51–83 years (mean 64.5 years)

were included. The MD values of the visual fields ranged from -10.0 to +0.3 dB (mean -5.8 dB) for the patients with glaucoma and from -3.5 to +2.4 dB (mean +0.2 dB) for the healthy subjects.

Visual field defects, defined as either arcuate shaped in superior or inferior part of the field, or nasally, respectively, centrally placed defects, as seen in grey-scale maps, were clearly visible in 93% of visual field printouts from the patients with glaucoma. GHT classified 84% of the glaucoma visual fields as outside normal limits, 9% as borderline and 7% as within normal limits. Using cluster analysis for glaucoma, defined as minimum three or more non-edge adjacent depressed test points on $p < 0.05$ significance level with at least one test point at the $p < 0.01$ significance level as seen in the pattern deviation probability maps (Katz et al. 1991; Anderson 1999), 95% of the glaucoma fields were outside normal limits.

All 165 visual fields were assessed by thirty clinicians: four glaucoma experts, eight general ophthalmologists, 10 physicians with other subspecialties and eight ophthalmology residents.

As shown in Table 1, the ANN had a sensitivity of 93%, a value that was significantly better than the 83% noted for the average physician ($p < 0.001$). Sensitivity was highest for the ANN followed by the glaucoma experts, and it was lowest for the other subspecialists. Specificity was very similar for the different groups of physicians and the ANN.

The average classification error score was 1.80 (min. 0.88; max. 2.68) for the physicians and 1.50 for the ANN ($p = 0.001$). The glaucoma experts constituted the most successful clinician subgroup, with an average error score of 1.48. The scores for the other subgroups were as follows (Fig. 2): 1.60 for the general ophthalm-

Table 1. Sensitivity and specificity of the artificial neural network (ANN) and subjective assessors with varying experience of glaucoma.

	Sensitivity mean, % (min %, max %)	Specificity mean, % (min %, max %)
ANN	93*	91
Average physician, <i>n</i> = 30	83 (62, 96)	90 (59, 100)
Glaucoma experts, <i>n</i> = 4	87 (79, 93)	91 (86, 97)
General ophthalmologists, <i>n</i> = 8	85 (78, 95)	88 (59, 96)
Other subspecialists, <i>n</i> = 10	82 (62, 92)	92 (77, 100)
Residents, <i>n</i> = 8	83 (75, 96)	89 (59, 97)

* Significantly better compared to the average physician, $p < 0.001$.

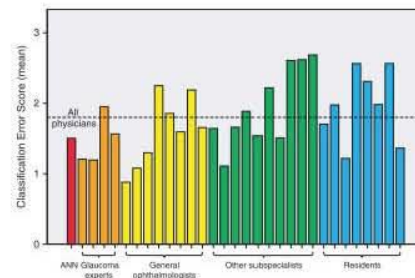


Fig. 2. Distribution of classification error scores for the artificial neural network (ANN) and all physicians. A lower score indicates both more correct and higher certainty in classification compared with a higher score. The vertical line indicates the average physician's score. The classification error score of the ANN was within the best third of the clinicians' scores and in parity with the scores of the glaucoma experts.

Coding Density

Weight of Evidence - A

Validation Process

ANN Performance

Weight of Evidence - C

Number of Cases Examined



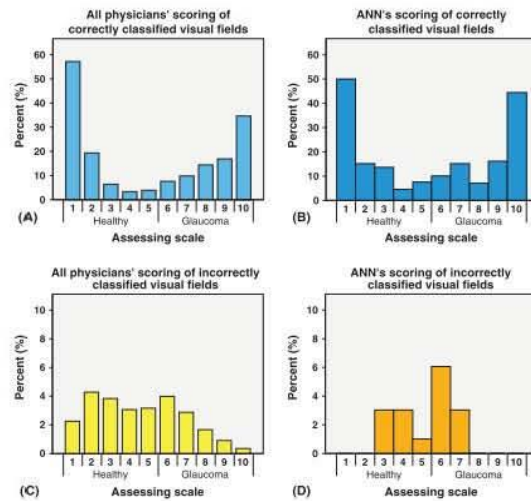


Fig. 3. Results of visual field scoring performed by the clinicians (A and C) and by the artificial neural network (ANN) (B and D). The scale indicates the certainty of assessments, with 1 representing healthy with absolute certainty and 10 absolute certainty of glaucoma. Patterns of the scoring of correctly classified fields (A and B) are similar for ANN and clinicians, with the highest frequencies at 1 and 10. Considering the incorrectly classified fields (C and D), the ANN (D) had scores closer to the borderline (i.e., 5 and 6) compared with the clinicians (C), who scored a few glaucomatous visual field tests as healthy with high certainty and a few healthy visual field tests as glaucomatous with high certainty.

mologists, 1.95 for other subspecialists and 1.96 for the residents.

The ANN scored 44% of the correctly classified visual fields of glaucoma patients with absolute certainty (score 10); the corresponding proportion for the average clinician was 35%. Also, 50% and 57% of the fields of healthy subjects were classified as healthy with absolute certainty (score 1) by the ANN and the average clinician, respectively (Fig. 3A,B). Considering the incorrectly classified fields, the clinicians' scores were distributed along the full scale from 1 to 10; 2% of the fields from glaucomatous eyes were scored as healthy with absolute certainty and 4% as almost certainly healthy. Fields that were incorrectly classified by the ANN had scores ranging from 3 to 7, that is, closer to a borderline classification (Fig. 3C,D). Thus, the ANN never made any incorrect classifications that were indicated as highly certain

(scores of 1 or 2 for glaucomatous eyes, or 9 or 10 for healthy eyes).

Discussion

The diagnostic accuracy of the SAP visual fields interpreted by the trained ANN was at least as good as those performed by clinicians who had full access to the Statpac interpretation tools. Also, the ANN offered better sensitivity, and its specificity was similar to that noted for the average physician.

The classification error score showing certainty of the assessments was at a level similar to the best third of the scores for all physicians. The networks classification error score was also similar to the scores of the glaucoma experts; however, this has to be interpreted cautiously because of the low number of participating glaucoma experts. Obviously, the certainty of an assessment is important when the

diagnosis of glaucoma is based on the visual field. The certainty of assessments can be estimated by using ANNs to interpret visual field test results, and this can be achieved in clinical practice without any costs in terms of time for either physicians or their patients. Of course, the network output should not be considered as the ultimate diagnostic truth. A final diagnosis is made by the health professional in the context of other clinical parameters than the visual field, for example, IOP, risk factors, optic disc findings, but our results seem to indicate that the network can perform the field assessment part on the same level as most clinicians.

To our knowledge, we are the first to suggest and test the use of ANNs to estimate certainty in diagnostic assessments in perimetry. Our ANN was not completely erroneous in any of its classifications of the investigated visual fields, that is, it did not incorrectly assess any healthy field as definitely glaucomatous or any glaucomatous field as healthy. However, 193 of the physicians' classifications were entirely inaccurate, that is, they assessed glaucomatous visual fields as healthy with high certainty (scores of 1 or 2 on the assessment scale), and 25 fields of healthy subjects were incorrectly classified as glaucoma with high certainty (scores of 9 or 10). These observations suggest that the ANN can be used to classify visual fields as glaucomatous or healthy and also to indicate the certainty of the classification, that is, whether it is either very certain or more uncertain and thus closer to borderline.

Only a few previous studies have examined machine-learning classifiers in comparison with subjective assessment of visual field tests (Goldbaum et al. 1994, 2002), and to our knowledge, no researchers have evaluated a fully trained machine-learning classifier using an independent sample. In the earlier investigations by Goldbaum and colleagues, the subjective assessors were glaucoma experts who had only limited or no access to the Statpac interpretation tool (Goldbaum et al. 1994, 2002). We believe that giving clinicians access to contemporary standard tools for interpretation of visual field test results should provide a good benchmark for assessing the performance of our ANN. We

included not only glaucoma experts but also several other categories of physicians in our study, because in many cases, clinicians without expert training in glaucoma are required to manage patients with glaucoma.

In conclusion, our findings indicate that a well-designed and validated ANN can perform at least as well as physicians in classifying SAP test results as healthy or glaucomatous. By incorporating such a network into the standard armamentarium of computer-assisted analyses, it should be possible to provide a similar high standard in many settings of glaucoma care, including outpatient clinics run by ophthalmic nurses or optometrists.

Acknowledgements

We are grateful to all the participating physicians for their interest and collaboration in this project. The study was supported by Swedish Research Council grant K2011-63X-10426-19-3, the Herman Järnhardt Foundation, the Foundation for Visually Impaired in Former Malmöhus County and Crown Princess Margareta's Foundation for the Visually Impaired.

References

- Anderson DR (1999): Automated Static Perimetry. St Louis, MO, USA: Mosby, Inc.
- Asman P & Heijl A (1992): Glaucoma hemifield test. Automated visual field evaluation. *Arch Ophthalmol* **110**: 812-819.
- Bengtsson B & Heijl A (1999): Inter-subject variability and normal limits of the SITA Standard, SITA Fast, and the Humphrey Full Threshold computerized perimetry strategies. *SITA STATPAC*. *Acta Ophthalmol Scand* **77**: 125-129.
- Bengtsson B & Heijl A (2000): False-negative responses in glaucoma perimetry: indicators of patient performance or test reliability? *Invest Ophthalmol Vis Sci* **41**: 2201-2204.
- Bengtsson B, Bizios D & Heijl A (2005): Effects of input data on the performance of a neural network in distinguishing normal and glaucomatous visual fields. *Invest Ophthalmol Vis Sci* **46**: 3730-3736.
- Bizios D, Heijl A & Bengtsson B (2007): Trained artificial neural network for glaucoma diagnosis using visual field data: a comparison with conventional algorithms. *J Glaucoma* **16**: 20-28.
- Chan K, Lee TW, Sample PA, Goldbaum MH, Weinreb RN & Sejnowski TJ (2002): Comparison of machine learning and traditional classifiers in glaucoma diagnosis. *IEEE Trans Biomed Eng* **49**: 963-974.
- Flammer J, Drance SM, Jenni A & Bebie H (1983): JO and STATJO: programs for investigating the visual field with the Octopus automatic perimeter. *Can J Ophthalmol* **18**: 115-117.
- Frankhauser F, Spahr J & Bebie H (1977): Some aspects of the automation of perimetry. *Surv Ophthalmol* **22**: 131-141.
- Goldbaum MH, Sample PA, White H, Colt B, Raphaelian P, Fechtner RD & Weinreb RN (1994): Interpretation of automated perimetry for glaucoma by neural network. *Invest Ophthalmol Vis Sci* **35**: 3362-3373.
- Goldbaum MH, Sample PA, Chan K et al. (2002): Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Invest Ophthalmol Vis Sci* **43**: 162-169.
- Goldbaum MH, Sample PA, Zhang Z et al. (2005): Using unsupervised learning with independent component analysis to identify patterns of glaucomatous visual field defects. *Invest Ophthalmol Vis Sci* **46**: 3676-3683.
- Heijl A (1986): A package for the Statistical Analysis of Visual Fields. Amsterdam: Martinus Nijhoff/DR W. Junk Publishers.
- Heijl A, Lindgren G, Olsson J & Asman P (1989): Visual field interpretation with empiric probability maps. *Arch Ophthalmol* **107**: 204-208.
- Henson DB, Spenceley SE & Bull DR (1996): Spatial classification of glaucomatous visual field loss. *Br J Ophthalmol* **80**: 526-531.
- Katz J & Sommer A (1988): Reliability indexes of automated perimetric tests. *Arch Ophthalmol* **106**: 1252-1254.
- Katz J, Sommer A, Gaasterland DE & Anderson DR (1991): Comparison of analytic algorithms for detecting glaucomatous visual field loss. *Arch Ophthalmol* **109**: 1684-1689.
- Sample PA, Goldbaum MH, Chan K et al. (2002): Using machine learning classifiers to identify glaucomatous change earlier in standard visual fields. *Invest Ophthalmol Vis Sci* **43**: 2660-2665.
- Sample PA, Chan K, Boden C et al. (2004): Using unsupervised learning with variational bayesian mixture of factor analysis to identify patterns of glaucomatous visual field defects. *Invest Ophthalmol Vis Sci* **45**: 2596-2605.
- Tucker A, Vinciotti V, Liu X & Garway-Heath D (2005): A spatio-temporal Bayesian network classifier for understanding visual field deterioration. *Artif Intell Med* **34**: 163-177.

Received on October 4th, 2011,
Accepted on February 23rd, 2012.


Correspondence:
Sabina Andersson
Department of Clinical Sciences
Ophthalmology in Malmö
Skåne University Hospital
SE-205 02 Malmö
Sweden
Tel: + 46 40 332757
Fax: + 46 40 336212
Email: sabina.andersson@med.lu.se

Copyright of Acta Ophthalmologica (1755375X) is the property of Wiley-Blackwell and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Copyright Clearance Center

Weight of Evidence Analysis – McLaren, Chen, Nie, & Su (2009)

WEIGHT OF EVIDENCE ANALYSIS		WoE-A				WoE-B		WoE-C			WoE-D	Accept?
Study ID	Author/Date	# Cases	Validation	D/P Appl	SCORE	ANN Study	SCORE	ANN Detail	ANN Perf	SCORE	TOTAL	Yes/No
215	McLaren et al, 2009	Yes	Yes	Yes	3	Yes	3	Some	Yes	2	8	Yes



NIH Public Access
Author Manuscript
Acad Radiol. Author manuscript; available in PMC 2010 July 1.

Published in final edited form as:
Acad Radiol. 2009 July ; 16(7): 842-851. doi:10.1016/j.acra.2009.01.029.

PREDICTION OF MALIGNANT BREAST LESIONS FROM MRI FEATURES: A COMPARISON OF ARTIFICIAL NEURAL NETWORK AND LOGISTIC REGRESSION TECHNIQUES

Christine E. McLaren, Ph.D.^{1,2}, Wen-Pin Chen, M.D.², Ke Nie, M.S.³, and Min-Ying Su, Ph.D.³

¹Department of Epidemiology, University of California, Irvine
²Chao Family Comprehensive Cancer Center, University of California Irvine
³Tu & Yuen Center for Functional Onco-Imaging, University of California, Irvine

Abstract

Rationale and Objectives—Dynamic contrast enhanced MRI (DCE-MRI) is a clinical imaging modality for detection and diagnosis of breast lesions. Analytical methods were compared for diagnostic feature selection and performance of lesion classification to differentiate between malignant and benign lesions in patients.

Materials and Methods—The study included 43 malignant and 28 benign histologically-proven lesions. Eight morphological parameters, ten gray level co-occurrence matrices (GLCM) texture features, and fourteen Laws’ texture features were obtained using automated lesion segmentation and quantitative feature extraction. Artificial neural network (ANN) and logistic regression analysis were compared for selection of the best predictors of malignant lesions among the normalized features.

Results—Using ANN, the final four selected features were compactness, energy, homogeneity, and Law_LS, with area under the receiver operating characteristic curve (AUC) = 0.82, and accuracy = 0.76. The diagnostic performance of these 4-features computed on the basis of logistic regression yielded AUC = 0.80 (95% CI, 0.688 to 0.905), similar to that of ANN. The analysis also shows that the odds of a malignant lesion decreased by 48% (95% CI, 25% to 92%) for every increase of 1 SD in the Law_LS feature, adjusted for differences in compactness, energy, and homogeneity. Using logistic regression with z-score transformation, a model comprised of compactness, NRL entropy, and gray level sum average was selected, and it had the highest overall accuracy of 0.75 among all models, with AUC = 0.77 (95% CI, 0.660 to 0.880). When logistic modeling of transformations using the Box-Cox method was performed, the most parsimonious model with predictors, compactness and Law_LS, had an AUC of 0.79 (95% CI, 0.672 to 0.898).

Conclusion—The diagnostic performance of models selected by ANN and logistic regression was similar. The analytic methods were found to be roughly equivalent in terms of predictive ability when

© 2009 The Association of University Radiologists. Published by Elsevier Inc. All rights reserved.
 Address for Correspondence: Christine E. McLaren, Professor and Director of Biostatistics, Department of Epidemiology, College of Health Sciences, University of California, Irvine, 224 Irvine Hall, Irvine, California 92697-7550, Telephone: 949-824-4007, Facsimile: 949-824-4773, E-mail: emclaren@uci.edu.
Publisher’s Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

a small number of variables were chosen. The robust ANN methodology utilizes a sophisticated non-linear model, while logistic regression analysis provides insightful information to enhance interpretation of the model features.

INTRODUCTION

Dynamic contrast enhanced MRI (DCE-MRI) is a clinical imaging modality for detection and diagnosis of breast lesions. Computer algorithms have been employed for automated lesion segmentation (1–4) with statistical analysis techniques applied to select an optimal set of features to achieve the highest diagnostic accuracy (5). Breast MRI has demonstrated a high sensitivity (>95%) with specificity of approximately 67% reported by meta analysis (6–11). To address the challenge of accuracy and efficiency in interpretation of breast MRI, a computer-aided diagnosis (CAD) system that can automatically analyze lesion features to differentiate between malignant and benign lesions would be very useful.

The general approach of breast CAD to predict a dichotomous outcome, malignant versus benign, involves applying computer algorithms for tumor characterization and then developing models using techniques including linear discriminate analysis (12), logistic regression analysis (13), and artificial neural networks (ANN) (14) for classifying a lesion as either malignant or benign. Because of inherent differences in these techniques, it is of interest to compare the diagnostic performance of analysis methods. In a review of 28 studies comparing ANN with other statistical approaches, Sargent found that ANN outperformed regression analysis in 36% of cases, ANN was outperformed by regression analysis in 14% of cases, and that similar results were obtained in 50% of cases (15). The advantages of ANN for predicting a dichotomous outcome include a requirement of less formal statistical training, an ability to implicitly detect complex nonlinear relationships between dependent and predictor variables, consideration of all possible interactions between predictor variables, and the availability of multiple training algorithms. Disadvantages include the “black box” nature of the ANN procedure, a greater computational burden, the tendency for overfitting, and the empirical nature of model development (16). Logistic regression is superior for examining possible causal relationships between independent and dependent variables, and understanding the effect of predictors on outcome variables (16). In summary, there is no universal approach to select models for data classification; evaluation of tasks on a case-by-case basis is necessary.

To our knowledge, only one study has compared ANN and logistic regression for breast cancer diagnosis using sonograms by Song et al. to differentiate between 24 malignant and 30 benign lesions (17). No difference was found in the diagnostic performance of ANN and logistic regression when measured by the area under the ROC curve. We have developed a CAD that incorporates a clustering-based algorithm for lesion segmentation and derives a full panel of quantitative morphological and texture descriptors for lesion characterization (18). In that study the ANN was used to select diagnostic features. In the present work we investigated and compared the diagnostic performance using the ANN and logistic regression.

Two aims were studied. Firstly we applied ANN to select features and investigated the diagnostic performance of different feature sets (models) selected based on the morphology and texture of the lesion. While ANN can select features robustly, the non-linear diagnostic model makes the weighting of each individual feature not interpretable. The logistic regression was applied to gain more insightful understanding for how each selected feature can be interpreted. Secondly, we illustrated the use of logistic regression for feature selection. The following statistical procedures were applied sequentially: 1) transformation of feature values to induce normality, 2) consideration of strategies for model selection and validation, 3) assessment of correlation between features to reduce collinearity, 4) reduction of the number of variables in models to avoid overfitting, and 5) evaluation and comparison of the diagnostic

McLaren et al.

Page 3

performance of classifiers. The selected features were compared to those selected by ANN. Also the diagnostic performance achieved by each selected model was evaluated and compared to that achieved by ANN.

MATERIALS AND METHODS

Malignant and Benign Lesion Database

The database analyzed in this study included 43 malignant and 28 benign lesions, the same database as reported in a previous study (18). The 43 patients with malignant lesions were from 29 to 76 years old with mean (\pm SD) 48 \pm 9 yr and median 48 yr. The 28 patients with benign lesions were from 21 to 74 yr, with mean 45 \pm 7 yr. and median 45 yr. The MRI studies were performed on a Philips Eclipse 1.5T scanner (Cleveland, OH). The images were acquired using a T1-weighted 3D SPGR (RF-FAST) pulse sequence, with TR=8.1 ms, TE=4.0 ms, flip angle=20°, matrix size=256 \times 128, FOV (Field of View) varying between 32 and 38 cm. The lesion was identified based on contrast enhanced images at 1-min after injection of the MR contrast medium Omniscan® (0.1 mmol/kg body weight).

The lesion was segmented on contrast-enhanced images. The detailed procedures have been reported previously (18). The operator inspected the lesion presentations on MRI, and determined the beginning and the ending image slices that contain the lesion, then one square box was placed on the lesion on one image slice to indicate the lesion location. With these inputs the 3-dimensional boundary of the lesion (or, lesion ROI- Region Of Interest) was obtained using an automated computer program based on the fuzzy c-means (FCM) clustering algorithm (1). After the 3-dimensional ROI for one lesion was defined, computer algorithms were used to extract quantitative features representing the morphological and texture properties of this lesion.

Eight morphological parameters were obtained for each tumor, including: volume, surface area, compactness, normalized radial length (NRL) mean, sphericity, NRL entropy, NRL ratio, and roughness, were obtained. Texture is a repeating pattern of local variations in image intensity, and is characterized by the spatial distribution of intensity levels in a neighborhood. Therefore, the texture features analyzed within the contrast enhanced lesion represent the distribution of enhancements. Ten GLCM texture features (energy, maximum probability, contrast, homogeneity, entropy, correlation, sum average, sum variance, difference average, and difference variance), as defined by Haralick and colleagues, were obtained for each lesion (19).

In addition, commonly used Laws' texture features were also obtained. The Laws' features were computed by first applying small convolution kernels to a digital image and then performing a nonlinear windowing operation. The 2D convolution kernels were generated from a set of 1D convolution kernels: L5=[1 4 6 4 1], E5=[-1 -2 0 2 1], S5=[-1 0 2 0 -1], R5=[1 -4 6 -4 1], W5=[-1 2 0 -2 -1]. These mnemonics stood for Level, Edge, Spot, Ripple and Wave. 2D convolution kernels such as Law_LE were obtained by convolving a vertical L5 kernel with a horizontal E5 kernel. In total, a set of 14 texture features that were rotationally invariant were obtained (Law_LE, Law_LS, Law_LW, Law_LR, Law_FE, Law_ES, Law_EW, Law_ER, Law_SS, Law_SW, Law_SR, Law_WW, Law_WR, and Law_RR) (20).

Artificial Neural Network (ANN) for diagnostic feature selection

A three-layer back-propagation neural network, known as multi-layer perceptrons (MLP) artificial neural network (ANN) was utilized to obtain optimal classifiers. The three-layer topology has an input layer, one hidden layer, and an output layer. The number of nodes in the input corresponds to the number of input variables. The output layer contains one node with

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Coding Density

Weight of Evidence - A

Number of Cases Examined

Weight of Evidence - C

Provides ANN Details

McLaren et al.

Page 4

values from 0 to 1 indicating level of malignancy, where 0 means absolutely benign and 1 means absolutely malignant. The number of hidden nodes is usually determined by a number of trial-and-error runs. Different neural network architectures with hidden nodes from 2 to 31 were tested. A stochastic gradient descent with the mean squared error function was used as the learning algorithm. The optimal architecture was chosen as the one for which the validation error was the lowest. With the determined number of hidden nodes, both the learning rate (0.1–0.7) and the momentum coefficient (0.2–0.8) were varied during network training to ensure a high probability of global network convergence. For training, criteria for convergence was met with a root mean squared error less than or equal to 0.0001 or a maximum of 1000 iterations. With the determined number of hidden nodes, both the learning rate, (0.1–0.7), and the momentum coefficient, (0.2–0.8), were varied during network training to ensure a high probability of global network convergence.

After the topology was chosen, feature selection was performed to find potential predictors within each set of morphology, GLCM texture, and Laws' texture features. The features were selected using the LNKnet (<http://www.ll.mit.edu/IST/lnknet/>) package in order to identify those yielding maximum discrimination capability thus achieving the optimal diagnostic performance. Each feature from 71 lesions was transformed to have zero mean and unit variance before training; after determination of the mean and standard deviation (STD) for a particular feature with $n=71$, the transformation for the i^{th} value of a feature is $z_i = (\text{feature value} - \text{mean})/\text{STD}$. This transformation is known as a z-score.

A forward search strategy was applied to find the optimal feature subset, which was obtained when the trained classifier produced the least error rate. The k -fold stratified cross-validation technique with $k=4$ was applied to test the performance of the generated classifiers. Weights and bias of the neural network were determined by a two-phase training procedure. The first phase had 30 iterations of back propagation, and the second phase had a longer run of conjugated gradient descent to ensure full convergence. The logistic sigmoid function was used to interpret the output variation in terms of probability of class membership within the range (0 to 1). To control for overfitting, the potential feature set was limited to no more than 4 in each category which had the least error. After the features from each of three categories (morphology, GLCM, and Laws') were individually obtained (Models A, B, and C respectively), all selected predictors were considered in a combined model (Model D), and a final model, restricted to contain no more than four features (Model E), was selected using ANN. Specifically, these models consisted of Model A with three features selected from eight morphology features, Model B with three features selected from 10 GMLM features, and model C with one selected from 14 Laws' features, Model D consisting of the combined seven selected features, and Model E consisting of four features selected from the combined seven features.

Logistic Regression Model

For this study the binary response variable of interest is the presence of a malignant breast lesion, coded 1=yes and 0=no. Note that the mean of the response values for a sample of lesions is the same as the proportion of lesions that are malignant. A logistic regression model that predicts a transformation of the response variable is used, $\text{logit}(p)$. Let p = probability that a lesion is malignant and $1-p$ = probability that a lesion is benign. Then the log odds is defined as

$$\text{logit}(p) = \log_e \left(\frac{p}{1-p} \right) \quad (1)$$

The logistic regression model enables prediction of the probability of a malignant breast lesion in relation to m lesion features, $x_1, x_2, x_3, \dots, x_m$ from an equation of the form

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Coding Density

Weight of Evidence - A

Validation Process

McLaren et al.

Page 5

$$\log_e \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_m x_m \quad (2)$$

where β_0 is the intercept term, and $\beta_1, \beta_2, \beta_3, \dots, \beta_m$, are the coefficients in the model associated with the m lesion features. It is assumed that the lesion features are linearly related to the log odds of the response. For a continuous lesion feature, the odds ratio for malignant versus benign lesions can be estimated as the exponentiation of the associated coefficient in the model (21).

Evaluation of ANN models using Logistic Regression

To address the first aim of the study, the selected diagnostic features in the five final models (A-E) by ANN were analyzed using logistic regression. Lesions were classified as benign or malignant on the basis of increasing threshold probabilities of malignancy. The ROC curve was constructed from the full range of probability thresholds with corresponding data points (sensitivity, 1-specificity) and the AUC was calculated. A nonparametric approach was used to compute the 95% confidence interval for the AUC (22,23). The classification of lesions on the basis of probability of being malignant > 0.5 was used to compare results for logistic regression with ANN for accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Logistic Regression Analysis for Diagnostic Feature Selection

To address the second aim of the study, logistic regression was used for initial selection of predictive models from feature z -scores. Descriptive statistics were calculated for features obtained from 71 lesions. The k -fold stratified cross-validation was applied with $k=4$. Forty-three malignant and twenty-eight benign lesions were separately and randomly assigned into four sub-cohorts so that each sub-cohort contained approximately the same proportion of malignant cases as the original cohort. To build predictive models, four sets of any three sub-cohorts were combined and used as a training cohort, with the remaining sub-cohort used as the corresponding validating cohort. These were labeled training cohorts 1–4 with corresponding validating cohorts 1–4.

The analysis method is now summarized. For each training set, logistic regression analysis was applied to examine the association between each individual feature and the dichotomous outcome, malignant lesion versus benign lesion. For example, univariate logistic regression models were formed, one for each of the eight morphological features. For each model, the likelihood ratio statistic was computed and the Bonferroni-Holm multiple comparisons method was applied to indicate statistically significant features at a 5% experimentwise significance level. These procedures were repeated for the 10 GLCM texture features and for the 14 Laws' texture features and the statistically significant features for each group were retained.

Next, for each training set, stepwise logistic model selection was applied to each set of features. For example, stepwise logistic regression was applied to the eight morphological features as predictors. The Score chi-square statistics were calculated and used to determine addition and removal of variables at the 0.05 significance level. The stepwise procedure was repeated for the 10 GLCM texture features and for the 14 Laws' texture features. Due to the small sample size of the cohorts, at most two predictors were retained from each set of features.

Pearson's correlation coefficient was computed for the pairs of predictors retained on the basis of univariate and stepwise modeling. Multivariate models were then built considering the combined morphology, GLCM texture and Laws' texture features that did not have statistically significant pairwise correlation. For each multivariate model, the variance inflation factor

McLaren et al.

Page 6

(24) for features in a model were examined with the objective of minimizing multicollinearity. No effect of multicollinearity is indicated by a VIF of zero. A common rule of thumb is that if the variance inflation is >5 , then multicollinearity is high.

For each training set, a list of models was generated in ascending order of estimated AUC. The *c* index of concordance statistic was calculated to compare training datasets with regard to the receiver operating characteristic area (25,26) and the 95% confidence interval of the average of concordance statistics was obtained for each model. Moreover, the Hosmer-Lemeshow goodness-of-fit test was performed for each model (26,27). The models selected from each training cohort were applied to the corresponding validating sub-cohort.

Finally, the models with the greatest AUC for each validating sub-cohort were applied to the entire cohort. For each of these models, the accuracy, sensitivity, specificity, PPV, and NPV were calculated.

The odds of a lesion to be malignant was computed for every increase of 1 SD in a feature, adjusted for differences in the other features in the model. For the comparison, the same analysis method was used after Box-Cox transformations were applied to induce normality, as needed.

RESULTS

Artificial Neural Network Diagnostic Feature Selection and Evaluation

Before ANN training, features were transformed to z-scores based on the mean and standard deviation of the entire set of 71 lesions. A three layer ANN with 5 hidden nodes in the hidden layer was chosen after a number of trial-and-error runs. The diagnostic measures used to assess differentiation between 43 malignant and 28 benign lesions are summarized in Table 1. Considering the eight morphology features, the classifier selected by ANN included three parameters: lesion volume, NRL entropy, and compactness (Model A). Using these three features for ROC analysis of the entire cohort of 71 lesions, an AUC of 0.80 and accuracy of 0.77 was achieved. Considering the ten GLCM texture features, the selected parameters were: gray level energy, gray level sum average, and homogeneity (Model B), with an AUC of 0.81 and accuracy of 0.73. From fourteen Laws' energy features, the best classifier contained only one parameter, Law_LS (Model C), with an AUC of 0.70 in the ROC analysis and accuracy of 0.65. Applying ANN to the combined set of seven features, including compactness, lesion volume, and NRL entropy, energy, gray level sum average, homogeneity, and the Law_LS parameter (Model D), the resulting AUC and accuracy were improved to 0.87 and 0.79, respectively, with positive and negative predictive values of 0.83 and 0.72. When these seven features were considered by a further ANN selection process, the final four selected features were: compactness, energy, homogeneity, and Law_LS, (Model E) and achieved an AUC of 0.82, accuracy of 0.76, PPV of 0.78, and NPV of 0.72.

Logistic regression was applied to each of the models selected by ANN for the cohort of 71 lesions, classifying a lesion as malignant if the threshold probability of malignancy was >0.5 . The diagnostic measures are also displayed in Table 1 for comparison with the ANN results. The values for diagnostic measures are similar to those recorded from the ANN technique. For example, the 3-feature morphology model of volume, NRL entropy, and compactness (Table 1, model A) has an AUC of 0.80 recorded by the ANN technique, and the same AUC of 0.80 (95% CI, 0.686 to 0.912) computed using logistic regression. The 95% confidence interval for the AUC from logistic regression contains the 0.80 value recorded from ANN. Diagnostic accuracy of 0.77, recorded by ANN, was higher than the value of 0.76 computed from logistic regression. Similarly, the final 4-feature model of compactness, energy, homogeneity, and Law_LS, (Table 1, Model E), had an AUC of 0.82 recorded by the ANN technique, as compared to 0.80 (95% CI, 0.688 to 0.905) computed on the basis of logistic regression. The 95%

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Coding Density

ANN Performance
Weight of Evidence - C

McLaren et al.

Page 7

confidence interval for AUC contains the computed value of 0.82 from ANN. Figure 1 displays the ROC curves for the 4-feature model (E) analyzed using the ANN and logistic regression. The diagnostic accuracy of 0.76 recorded by ANN was slightly higher than 0.72 computed by logistic regression.

For the outcome of presence of a malignant lesion and modeling of z-scores from feature predictors with $n=71$, the logistic regression equation was

$$\text{logit}(p) = 0.69 + 1.69 \text{ Compactness} - 0.61 \text{ Energy} + 0.24 \text{ Homogeneity} - 0.73 \text{ Law_LS} \quad (3)$$

Law_LS was the most predictive variable in the model (Wald statistic, $P=0.028$). The P-value of the Hosmer-Lemeshow test was 0.76 indicating a reasonably good fit of the model to the data. However, the adjusted R^2 for the model was 0.30 which indicates that only 30% of the total variation in values could be explained by the predictor features in the model (28). The estimated odds ratio for Law_LS was calculated as the exponentiation of the associated estimated coefficient, $\exp(-0.73) = 0.48$. Thus the odds of a malignant lesion decreased by 48% (95% CI, 25% to 92%) for every increase of 1 SD in the Law_LS feature, adjusted for differences in compactness, energy, and homogeneity. For example, comparing a lesion with Law_LS of 0.19, one standard deviation above the mean of 0.13 for the set of 71 lesions, to a lesion Law_LS of 0.25, two standard deviations above the mean, the odds of the second lesion being malignant are about one-half of that of the first lesion, assuming that they have similar values for compactness, energy, and homogeneity.

Diagnostic Feature Selection using Logistic Regression Analysis

To address the second aim of the study, we applied logistic regression to z-scores for morphology, GLCM texture, and Laws' texture features and followed the modeling strategy described in Materials and Methods. Within training sets, features preselected for consideration by multivariate models had no statistically significant correlation between pairs of features ($P > 0.13$ for all). Table 2 displays the best single model selected from each training cohort. For each training cohort, the model chosen had the highest AUC from among those applied to the corresponding validation cohort. For a particular model, the diagnostic values presented in the table were calculated when the model was used to reclassify the full data set of 71 lesions on the basis of probability of being benign or malignant. Each of the four models (Table 2, training cohort 1–4) contained two morphological parameters, compactness and NRL Entropy. In addition, models chosen from training cohort 1 and 2 contained a GLCM texture feature. The model selected from training cohort 1 included energy; while the model from training cohort 2 included gray level sum average. Identical models were chosen from training cohorts 3 and 4 and contained the Laws' texture feature, Law_LW. When the chosen models were applied to the full data set of 71 lesions, all models had P-values for the likelihood ratio test less than 0.004 indicating the joint significance of all predictor features with respect to the outcome. Values for AUC varied from 0.75 for the model from training cohort 1 (compactness, NRL entropy, energy) to 0.80 for the model selected from training cohorts 3 and 4 (compactness, NRL entropy, Law_LW). However, the 95% confidence intervals for the AUC overlap, indicating no significant difference between AUC values. Estimated accuracy of models varied from 72% to 75%. The model chosen from training cohort 2 had the highest overall accuracy, correctly classifying 75% of lesions, and an AUC of 0.77 (95% CI, 0.660 to 0.880). This model was comprised of three features including compactness, NRL entropy, and gray level sum average. The model had high sensitivity, correctly identifying 91% of malignant lesions and moderate specificity, correctly identifying 50% of benign lesions. While the model selected from training cohort 1 (compactness, NRL entropy, and energy) had the highest sensitivity and correctly identified 95% of malignant cases, the specificity was only 39%.

McLaren et al.

Page 8

Figure 2 displays the ROC curve, illustrated with a solid line, for the 3-feature model selected from training cohort 2 with the highest overall accuracy among validating cohorts, correctly classifying 75% of the 71 lesions. The model was comprised of two morphological parameters, compactness and NRL entropy, and one GLCM feature, gray level sum average (Table 2, Figure 2). Applied to the entire cohort of 71 lesions, there was no statistically significant correlation between pairs of variables ($P>0.10$ for all). A good fit of the model to the data was demonstrated (Hosmer and Lemeshow goodness-of-fit statistic, $P=0.88$) and a high concordance statistic (0.78; 95% CI, 0.755 to 0.813) implied a good ability to predict the lesion type. As expected, the average variance inflation factor (VIF) for the three features in the model was 1.02 indicating low effect of multicollinearity on the variance of model coefficients. The strongest predictor in the model was gray level sum average. The odds of a malignant lesion increased by 2.0 (95% CI, 1.14 to 3.68) for every increase of 1 SD in the texture feature, gray level sum average, adjusted for differences in morphology features, compactness and NRL entropy. The logistic regression equation was

$$\text{logit}(p) = 0.66 + 1.28 \text{ Compactness} - 0.66 \text{ NRL Entropy} + 0.72 \text{ Gray Level Sum Average} \quad (4)$$

Gray level sum average was the most predictive variable in the model (Wald statistic, $P=0.016$). For this variable, the estimated odd ratio was $\exp(0.72) = 2.1$ (95% CI, 1.14 to 3.68), thus the odds of a malignant lesion increased by 2.1 for every increase of 1 SD in the gray level sum average feature, adjusted for differences in compactness and NRL entropy. For example, comparing a lesion with gray level sum average of 27.6, the mean for the 71 lesions, to a lesion with feature value of 34.2, one standard deviation above the mean, the odds of the second lesion being malignant are over twice that of the first lesion, assuming that they have similar values for compactness and NRL entropy.

We compared results of modeling the feature z-scores to those obtained after applying the Box-Cox method of selecting the most appropriate transformation to induce normality. A two-feature model of \ln compactness and Box-Cox transformed Law_LS had the highest accuracy of 0.79, with estimated AUC of 0.79 (95% CI, 0.672 to 0.898), sensitivity 0.88 and specificity 0.64. Figure 2 displays the ROC curves for the 2-feature model. The likelihood ratio chi-square for the model was statistically significant ($P<0.0001$). As expected, the correlation between \ln compactness and transformed Law_LS was not statistically significant ($r=0.196$, $P=0.10$) and a good fit of the model to the data was demonstrated (Hosmer and Lemeshow Goodness-of-Fit, $P=0.66$) with a concordance statistic 0.79 (95% C.I., 0.724 to 0.855). The average variance inflation factor was 1.04, indicating a low effect of multicollinearity on the variance of model coefficients. In the model, both features added information to the model (Wald statistic, $P<0.007$ for each). For the outcome of presence of a malignant lesion, the logistic regression equation was

$$\text{logit}(p) = -5.650 + 0.666 (\ln \text{ Compactness}) - 266.50 ((\text{Law_LS}^{(0.1)} - 1) / (1 / (0.1))) \quad (5)$$

For \ln compactness, the estimated odds ratio was $\exp(0.666) = 1.9$. This indicates that the odds of a malignant lesion increased by 1.9 (95% CI, 1.20 to 3.56) for every increase of 1 unit in the natural log of the morphology feature, compactness, adjusted for differences in the Laws' feature, Law_LS. For example, comparing a lesion with \ln compactness of 3.0, one standard deviation above the mean of 1.6 for the set of 71 lesions, to that of a lesion with \ln compactness of 4.5, two standard deviations above the mean, the second lesion would be nearly twice as likely to be malignant as first lesion, assuming equal values for Law_LS.

DISCUSSION

The primary aim of the study was to compare ANN and logistic regression analysis for lesion classification to differentiate between malignant and benign breast lesions in patients. Using our dataset of 71 lesions, the ANN procedure was applied to select the best classifiers for morphology and texture (GLCM and Laws') category features. The three selected morphology features (volume, NRL entropy, compactness) achieved a moderate AUC of 0.80 and estimated accuracy of 0.77. The three selected GLCM features (energy, gray level sum average, and homogeneity) achieved higher AUC (0.81) and estimated accuracy (0.73). Only one Laws' feature (Law_LS) was selected, and achieved lower AUC (0.70) and lower accuracy (0.65). When all seven features were combined the model achieved an improved AUC of 0.87 and estimated accuracy of 0.79. Submitting these seven features into another ANN selection, resulted in selection of four features, one morphology feature (compactness), two GLCM features (energy, homogeneity), and one Laws' texture feature (Law_LS) with an AUC of 0.82 and accuracy of 0.76. These results demonstrate that it is possible to use ANN to select the best combined indicators to predict tumor malignancy.

For the 4-feature model (Table 1, Model E), logistic regression analysis revealed that Law_LS was the most predictive variable in the model (Wald statistic, $P=0.028$). The estimated odds ratio for Law_LS was calculated as $\exp(-0.73) = 0.48$. Thus the odds of a malignant lesion decreased by 48% (95% CI, 25% to 92%) for every increase of 1 SD in the Law_LS feature, adjusted for differences in compactness, energy, and homogeneity. This result provides an illustration of the complementary manner in which ANN and logistic regression can be used (15). While ANN is more robust in that features are selected with minimal intellectual judgment of the operator, logistic regression can provide more insight and understanding into the relationship between selected features and the outcome.

Artificial neural networks have been used elsewhere in clinical data modeling, and similar results with that of regression modeling techniques were demonstrated (29–33). Compared to logistic regression modeling, ANN was found to have higher prediction rates in complex and non-linear relationships among a large number of variables; however even when the difference was significant (due to very large sample size) the improved performance was only marginal. Nilsson et al. used data from 18,362 patients undergoing cardiac surgery to predict the operative mortality. ANN selected 34 of the total 72 risk variables as relevant for mortality prediction. The area under ROC curve for ANN (0.81) was larger than that of the logistic regression model (0.79, $P=0.0001$) with the same 34 top-ranked risk variables (29). Delen and colleagues acquired a large dataset (202,932 cases with 17 variables) to predict five-year breast cancer survival using 10-fold cross-validation. The results indicated that when all variables were used, the ANN estimated area under the ROC curve with 0.91, compared to 0.89 computed by logistic regression (31). Lundin et al. tried to predict five-year breast cancer survival using data from 951 breast cancer patients. Using eight input variables, the ANN and logistic regression models achieved similar values for AUC of 0.901 and 0.897 respectively (33). Jaimes et al. (32) and Clermont et al. (30) also found that both ANN and logistic regression have similar performance when considering a small number of variables.

Advantages of neural network analysis are that few prior assumptions or knowledge about data distributions are required, so knowledge about complex variable transformations is not needed before training, and the search for the optimal diagnostic classifier involves minimal user input. Another advantage is that ANN has the capacity to model complex nonlinear relationships between independent and predictor variables, allowing the inclusion of a large number of variables. A disadvantage of ANN is the long training process and requirement of an experienced operator to determine the optimal network topology. The major factor that needs to be experimentally determined is the number of hidden layer nodes. If too few hidden nodes

are used, proper training is impeded. If too many are used, the neural network is over-trained. In our study the number of hidden nodes was determined by a number of trial-and-error runs. Another limitation of the ANN technique is the poor interpretability of selected models. Neither standardized coefficients nor odd ratios corresponding to selected variable can be calculated and presented as in regression models. Logistic regression can be applied in a complementary manner to provide this information, thus overcoming the problem. Furthermore, the technique can be used for hypothesis testing regarding univariate and multivariate associations between predictor variables and the outcome of interest (15) and to enhance understanding and interpretation of the effect of predictor variables on the response (16). The logistic regression analysis may be preferred to ANN due to improved interpretation of individual predictors.

Both ANN and logistic regression are subject to issues of overfitting, assessing model convergence, and collinearity that affect the generalization of results (15). Methods to avoid overfitting include cross-validation as applied in our study (34). Cross-validation serves to check internal validity (reproducibility) (35). Leave-one-out cross validation, although almost unbiased, may have high variance leading to unreliable estimates (36). Kohavi studied cross-validation for accuracy of estimation and model selection and recommends *k*-fold stratified cross validation for model selection on the basis of stability of predictions and accuracy, when compared to leave-one-out cross validation (34,37). As recommended, in this study we use 4-fold cross validation for ANN and logistic regression modeling. Finally, it is important to note that ANN does not strictly check collinearity among features during the selection process and collinearity can affect the variance of model estimates. Since the potential for collinearity among the features is not specifically taken into account, this can lead to stability problems (38–40).

The second aim of the study was to illustrate the use of logistic regression for feature selection. We examined differences between methods of standardizing features and addressed the issue of potential collinearity by incorporating statistical testing for correlation between variables to pre-select variables for further modeling. As illustrated in Table 2, the logistic regression model of feature *z*-scores with the highest estimated accuracy of 0.75 and AUC (0.77; 95% CI, 0.660 to 0.880) included compactness, NRL entropy, and gray level sum average (Figure 2). The model of Box-Cox transformed values with the highest accuracy of 0.79 contained two features, In compactness and Box-Cox transformed Law_LS, with estimated AUC of 0.79 (95% CI, 0.672 to 0.898), sensitivity 0.88 and specificity 0.64 (Figure 2). The results suggest that the diagnostic performance of the models selected by logistic regression was comparable to that of ANN; also that the method of standardization may improve on model selection.

Regarding the generalizability of our results, although we analyzed a relatively small dataset, the selected features have been well accepted and commonly used in the development of breast CAD systems. Using this same dataset, we previously attempted to establish the association between the extracted quantitative features and the lesion phenotype appearance on MRI as described in the BI-RADS breast MRI lexicon (18). For example, the compactness morphological parameter is strongly linked to the shape and margin of the lesion, and the GLCM texture features are associated with the degree of the enhancement and the heterogeneous enhancement patterns within the lesion. These BI-RADS descriptors are well-established diagnostic features. Therefore, although the generalization of the selected quantitative features need validation using independent datasets, they are generalizable in the sense that they are closely related to visual diagnostic features commonly used by radiologists.

In summary, we have shown that the diagnostic performance of models selected by ANN and logistic regression was similar and the analytic methods were found to be roughly equivalent in terms of predictive ability when a small number of variables were chosen. We have emphasized interpretation of the predictors in the model and illustrated comparison of lesion

McLaren et al.

Page 11

features in terms of the odds of a lesion being malignant, enhancing the usefulness of the logistic regression modeling. The ANN methodology is more robust (i.e. it does not require a high level of operator judgment), and it utilizes a sophisticated non-linear model to achieve a high diagnostic performance. On the other hand, logistic regression may generate many sets of models that yield similar diagnostic performance, and the operator will need to make intellectual judgments to select the best model(s). The modeling strategy provided in the present work requires statistical judgment and thus may be more difficult to implement in a large dataset that has a many variables compared to the “black box” ANN approach. Nonetheless, logistic regression analysis provides insightful information to enhance interpretation of the model features. Finally, many diagnostic models (feature sets) could be selected using ANN and logistic regression based on cross-validation within one dataset; and the ultimate diagnostic value of these models will have to be determined in an independent validation dataset.

Acknowledgments

This work was supported in part by NIH/NCI R01 CA90437 (O. Nalcioglu), CA121568 (M-Y Su), the California Breast Cancer Program grant #9WB-002 (M-Y Su), and the UC Irvine Cancer Center Support Grant No. 2P30CA062203-13S (F.L. Meyskens, Jr.)

REFERENCES

1. Chen W, Giger ML, Bick U. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Acad Radiol* 2006;13:63–72. [PubMed: 16399033]
2. Chen W, Giger ML, Bick U, et al. Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI. *Med Phys* 2006;33:2878–2887. [PubMed: 16964864]
3. Chen W, Giger ML, Lan L, et al. Computerized interpretation of breast MRI: investigation of enhancement-variance dynamics. *Med Phys* 2004;31:1076–1082. [PubMed: 15191295]
4. Liney GP, Sreenivas M, Gibbs P, et al. Breast lesion analysis of shape technique: semiautomated vs. manual morphological description. *J Magn Reson Imaging* 2006;23:493–498. [PubMed: 16523479]
5. Meinel LA, Stolpen AH, Berbaum KS, et al. Breast MRI lesion classification: improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system. *J Magn Reson Imaging* 2007;25:89–95. [PubMed: 17154399]
6. Esserman L, Hylton N, Yassa L, et al. Utility of magnetic resonance imaging in the management of breast cancer: evidence for improved preoperative staging. *J Clin Oncol* 1999;17:110–119. [PubMed: 10458224]
7. Fischer U, Kopka L, Grabbe E. Breast carcinoma: effect of preoperative contrast-enhanced MR imaging on the therapeutic approach. *Radiology* 1999;213:881–888. [PubMed: 10580970]
8. Mumtaz H, Hall-Craggs MA, Davidson T, et al. Staging of symptomatic primary breast cancer with MR imaging. *AJR Am J Roentgenol* 1997;169:417–424. [PubMed: 9242745]
9. Rieber A, Schirmeister H, Gabelmann A, et al. Pre-operative staging of invasive breast cancer with MR mammography and/or PET: boon or bunk? *Br J Radiol* 2002;75:789–798. [PubMed: 12381687]
10. Schelfout K, Van Goethem M, Kersschot E, et al. Contrast-enhanced MR imaging of breast lesions and effect on treatment. *Eur J Surg Oncol* 2004;30:501–507. [PubMed: 15135477]
11. Zhang Y, Fukatsu H, Naganawa S, et al. The role of contrast-enhanced MR mammography for determining candidates for breast conservation surgery. *Breast Cancer* 2002;9:231–239. [PubMed: 12185335]
12. Gilhuijs KG, Giger ML, Bick U. Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. *Med Phys* 1998;25:1647–1654. [PubMed: 9775369]
13. Chou YH, Tiu CM, Hung GS, et al. Stepwise logistic regression analysis of tumor contour features for breast ultrasound diagnosis. *Ultrasound Med Biol* 2001;27:1493–1498. [PubMed: 11750748]
14. Chan HP, Sahiner B, Petrick N, et al. Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network. *Phys Med Biol* 1997;42:549–567. [PubMed: 9080535]

McLaren et al.

Page 12

15. Sargent DJ. Comparison of artificial neural networks with other statistical approaches. *Cancer* 2001;91:1636–1642. [PubMed: 11309761]
16. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996;49:1225–1231. [PubMed: 8892489]
17. Song JH, Venkatesh SS, Conant EA, et al. Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. *Acad Radiol* 2005;12:487–495. [PubMed: 15831423]
18. Nie K, Chen J-H, Yu HJ, et al. Quantitative Analysis of Lesion Morphology and Texture Features for Diagnostic Prediction in Breast MRI. *Acad Radiol*. 2008 (in press).
19. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern* 1973;SMC-3:610–621.
20. Laws, KI. Society of Photo-Optical Instrumentation Engineers. San Diego, CA: Image processing for missile guidance; 1980. Rapid texture identification; p. 376–380.
21. Altman, DG. London: Chapman & Hall; 1991. Practical statistics for medical research.
22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–845. [PubMed: 3203132]
23. Puri, ML.; Sen, PK. New York: Wiley; 1971. Nonparametric Methods in Multivariate Analysis.
24. Neter, J.; Kutner, M.; Nachtsheim, C., et al. New York: WCB McGraw-Hill; 1996. Applied linear statistical models.
25. Harrell FE, Cadiff RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA* 1982;247:2543–2546. [PubMed: 7069920]
26. Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer. *J Natl Cancer Inst* 2007;99:715–726. [PubMed: 17470739]
27. Hosmer, DW.; Lemeshow, S. New York: John Wiley & Sons; 1989. Applied Logistic Regression.
28. Nagelkerke NJD. A Note on a General Definition of the Coefficient of Determination. *Biometrika* 1991;78:691–692.
29. Nilsson J, Ohlsson M, Thulin L, et al. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *J Thorac Cardiovasc Surg* 2006;132:12–19. [PubMed: 16798296]
30. Clermont G, Angus DC, DiRusso SM, et al. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit Care Med* 2001;29:291–296. [PubMed: 11246308]
31. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med* 2005;34:113–127. [PubMed: 15894176]
32. Jaimes F, Farbiarz J, Alvarez D, et al. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Crit Care* 2005;9:R150–R156. [PubMed: 15774048]
33. Lundin M, Lundin J, Burke HB, et al. Artificial neural networks applied to survival prediction in breast cancer. *Oncology* 1999;57:281–286. [PubMed: 10575312]
34. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. The Fourteenth International Joint Conference on Artificial Intelligence Morgan Kaufmann; San Mateo. 1995. p. 1137–1143.
35. Terrin N, Schmid CH, Griffith JL, et al. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *J Clin Epidemiol* 2003;56:721–729. [PubMed: 12954463]
36. Efron B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J Am Stat Assoc* 1983;78:316–331.
37. Breiman L, Spector P. Submodel Selection and Evaluation in Regression. The X-Random Case. *Int Stat Rev* 1992;60:291–319.
38. Martens, H.; Næs, T. Chichester: Wiley; 1989. Multivariate Calibration.
39. Næs T, Mevik B-H. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics* 2001;15:413–426.

McLaren et al.

Page 13

40. Weisberg, S. New York: Wiley; 1985. Applied Linear Regression.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Coding Density

McLaren et al.

Page 14

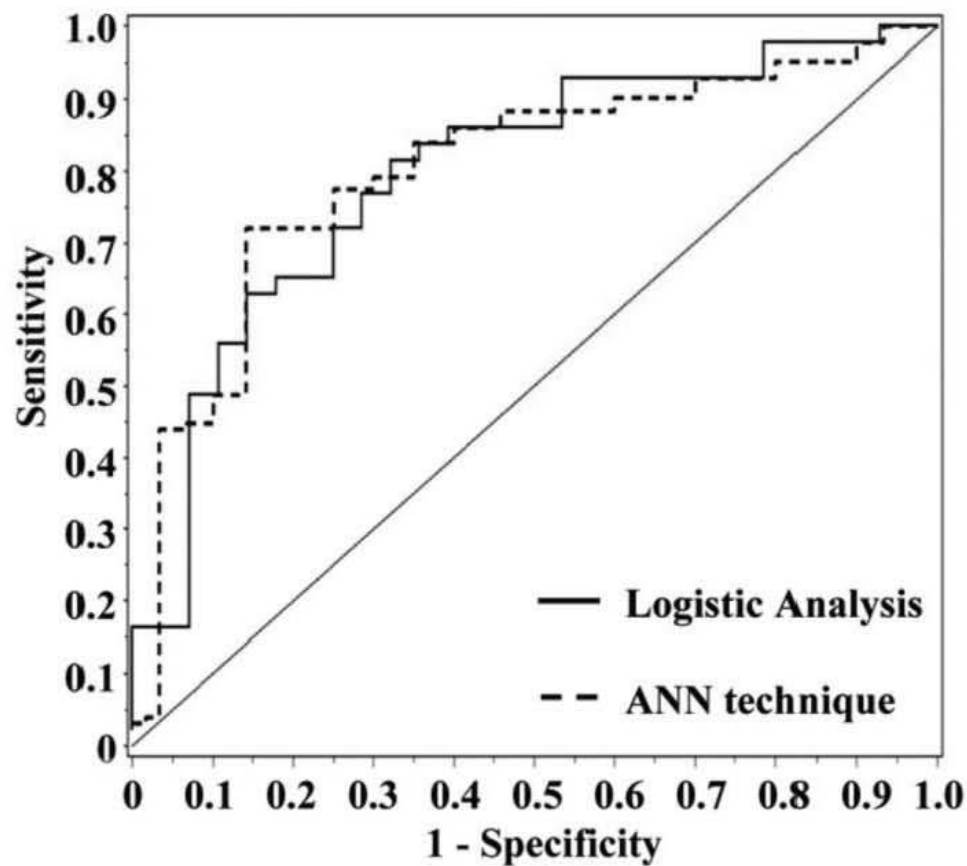


Figure 1. The dashed line represents the ROC curve for ANN modeling of z-scores (Table 1, Model E; Compactness, Energy, Homogeneity and Law_LS; AUC 0.82). The solid line represents the ROC curve for the four features as assessed by logistic regression (Table 1, Model E, AUC 0.80).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Coding Density

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

McLaren et al.

Page 15

Coding Density

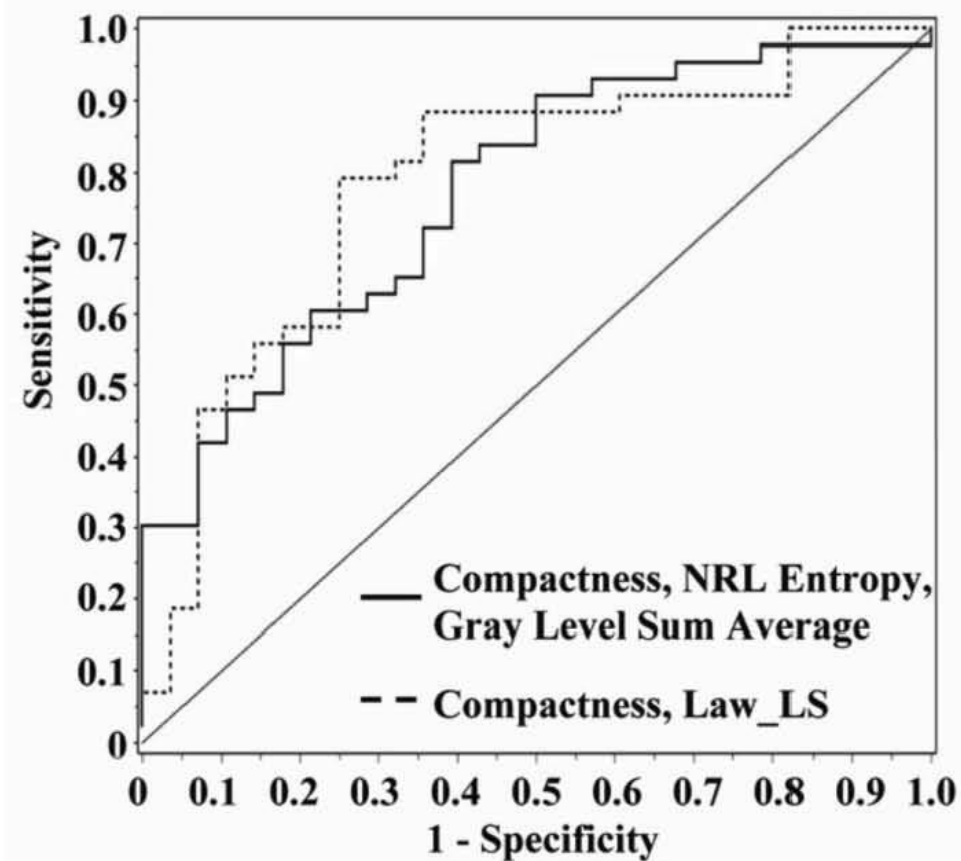


Figure 2.
The solid line represents the ROC curve for logistic regression modeling of z-scores for compactness, NRL entropy and gray level sum average (AUC 0.77; 95% CI, 0.660 to 0.880).
The dashed line represents logistic regression model of Box-Cox transformed values for compactness and Law_LS (AUC 0.79; 95% CI, 0.672 to 0.898).

McLaren et al.

Table 1

Diagnostic evaluation of models selected using the artificial neural network (ANN) technique. For each model the corresponding logistic regression equation also was applied to data from the full cohort (N = 71; Malignant = 43; Benign = 28).

Model	Imaging Descriptors ^d	Method	Accuracy/ (%)	Sensitivity ² (%)	Specificity ³ (%)	PPV ⁴	NPV ⁵	95% CI	
								Estimated Area under ROC	
				AUC					
A	Morphology (3 selected from 8) Volume, NRL Entropy, Compactness	ANN	77	88	61	0.78	0.77	0.80	
		Logistic Regression	76	93	50	0.74	0.82	0.80	0.686 0.912
B	GLCM (3 selected from 10) Energy, Gray Level Sum Average, Homogeneity	ANN	73	84	57	0.75	0.70	0.81	
		Logistic Regression	68	86	39	0.69	0.65	0.77	0.660 0.880
C	LAWs (only 1 selected from 14) Law_LS	ANN	65	84	36	0.67	0.59	0.70	
		Logistic Regression	65	86	32	0.66	0.60	0.70	0.573 0.819
D	Combining all 7 selected features Volume, NRL Entropy, Compactness, Energy, Gray Level Sum Average, Homogeneity, Law_LS	ANN	79	81	75	0.83	0.72	0.87	
		Logistic Regression	80	86	71	0.82	0.77	0.86	0.772 0.949
E	Final (4 selected from 7) Compactness, Energy, Homogeneity, Law_LS	ANN	76	84	64	0.78	0.72	0.82	
		Logistic Regression	72	86	50	0.73	0.70	0.80	0.688 0.905

¹Accuracy = (number of correctly identified cases / 71) × 100%

²Sensitivity = (number of correctly identified as malignant / 43) × 100%

³Specificity = (number of correctly identified as benign / 28) × 100%

⁴Positive Predictive Value (PPV) = number of lesions correctly identified as malignant / number of lesions identified as malignant

⁵Negative Predictive Value (NPV) = number of lesions correctly identified as benign / number of lesions identified as benign.

^dEach variable in the model was standardized by subtracting the mean and dividing by the standard deviation of data from 71 subjects.

Page 16

Coding Density

McLaren et al.

Table 2

Diagnostic evaluation of models selected using logistic regression of feature z-scores with 4-fold cross validation and applied to data from the full cohort (N = 71; Malignant = 43; Benign = 28).

Training Cohort	Model with highest AUC for corresponding validation cohort [#]	Likelihood ratio \ln^2	P-value	Accuracy ¹ (%)	Sensitivity ² (%)	Specificity ³ (%)	PPV ⁴	NPV ⁵	95% CI		
									Estimated Area under ROC ⁶	AUC	
1	Compactness, NRL Entropy, Energy	13.54	0.0036	73	95	39	0.71	0.85	0.75	0.626	0.879
2	Compactness, NRL Entropy, Gray Level Sum Average	17.86	0.0005	75	91	50	0.74	0.78	0.77	0.660	0.880
3,4	Compactness, NRL Entropy, Law_LW	19.16	0.0003	72	86	50	0.73	0.70	0.80	0.690	0.908

¹ Accuracy = (number of correctly identified cases / 71) × 100%.

² Sensitivity = (number of correctly identified as malignant / 43) × 100%

³ Specificity = (number of correctly identified as benign / 28) × 100%.

⁴ Positive Predictive Value (PPV) = number of lesions correctly identified as malignant / number of lesions identified as malignant

⁵ Negative Predictive Value (NPV) = number of lesions correctly identified as benign / number of lesions identified as benign.

⁶ Classification criteria: If the predicted value > 0.5, then the case was classified as malignant case by the model. If the predicted value < 0.5, then the case was classified as benign case by the model.

[#] Each variable in the model was standardized by subtracting the mean and dividing by standard deviation of values from 71 subjects.

Page 17

Coding Density


Research Question Analysis – McLaren, Chen, Nie, & Su (2009)

Study ID#	--- RQ#1 ---			--- RQ#2 ---						Comments	
	PERFORMANCE		OUTCOME	METHODOLOGY					APPLICATION		
	Performance Measure	Performance Value	Classification Successful?	Study Type	Primary Tool	Hybridized with...	Algorithm Specified	Comparative Analysis	ANN Best		Disease/Issue Application
215	AUROC	87.00%	Yes	Diagnostic	MLPBPNN	n/a	n/a	Yes	Equal	Breast Cancer	

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



NIH Public Access
Author Manuscript
Acad Radiol. Author manuscript; available in PMC 2010 July 1.

Published in final edited form as:
Acad Radiol. 2009 July; 16(7): 842-851. doi:10.1016/j.acra.2009.01.029.

PREDICTION OF MALIGNANT BREAST LESIONS FROM MRI FEATURES: A COMPARISON OF ARTIFICIAL NEURAL NETWORK AND LOGISTIC REGRESSION TECHNIQUES

Christine E. McLaren, Ph.D.^{1,2}, Wen-Pin Chen, M.D.², Ke Nie, M.S.³, and Min-Ying Su, Ph.D.³

¹Department of Epidemiology, University of California, Irvine
²Chao Family Comprehensive Cancer Center, University of California Irvine
³Tu & Yuen Center for Functional Onco-Imaging, University of California, Irvine

Abstract

Rationale and Objectives—Dynamic contrast enhanced MRI (DCE-MRI) is a clinical imaging modality for detection and diagnosis of breast lesions. Analytical methods were compared for diagnostic feature selection and performance of lesion classification to differentiate between malignant and benign lesions in patients.

Materials and Methods—The study included 43 malignant and 28 benign histologically-proven lesions. Eight morphological parameters, ten gray level co-occurrence matrices (GLCM) texture features, and fourteen Laws' texture features were obtained using automated lesion segmentation and quantitative feature extraction. Artificial neural network (ANN) and logistic regression analysis were compared for selection of the best predictors of malignant lesions among the normalized features.

Results—Using ANN, the final four selected features were compactness, energy, homogeneity, and Law_LS, with area under the receiver operating characteristic curve (AUC) = 0.82, and accuracy = 0.76. The diagnostic performance of these 4-features computed on the basis of logistic regression yielded AUC = 0.80 (95% CI, 0.688 to 0.905), similar to that of ANN. The analysis also shows that the odds of a malignant lesion decreased by 48% (95% CI, 25% to 92%) for every increase of 1 SD in the Law_LS feature, adjusted for differences in compactness, energy, and homogeneity. Using logistic regression with z-score transformation, a model comprised of compactness, NRL entropy, and gray level sum average was selected, and it had the highest overall accuracy of 0.75 among all models, with AUC = 0.77 (95% CI, 0.660 to 0.880). When logistic modeling of transformations using the Box-Cox method was performed, the most parsimonious model with predictors, compactness and Law_LS, had an AUC of 0.79 (95% CI, 0.672 to 0.898).

Conclusion—The diagnostic performance of models selected by ANN and logistic regression was similar. The analytic methods were found to be roughly equivalent in terms of predictive ability when

© 2009 The Association of University Radiologists. Published by Elsevier Inc. All rights reserved.
Address for Correspondence: Christine E. McLaren, Professor and Director of Biostatistics, Department of Epidemiology, College of Health Sciences, University of California, Irvine, 224 Irvine Hall, Irvine, California 92697-7550, Telephone: 949-824-4007, Facsimile: 949-824-4773, E-mail: cmclaren@uci.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Coding Density

a small number of variables were chosen. The robust ANN methodology utilizes a sophisticated non-linear model, while logistic regression analysis provides insightful information to enhance interpretation of the model features.

INTRODUCTION

Dynamic contrast enhanced MRI (DCE-MRI) is a clinical imaging modality for detection and diagnosis of breast lesions. Computer algorithms have been employed for automated lesion segmentation (1–4) with statistical analysis techniques applied to select an optimal set of features to achieve the highest diagnostic accuracy (5). Breast MRI has demonstrated a high sensitivity (>95%) with specificity of approximately 67% reported by meta analysis (6–11). To address the challenge of accuracy and efficiency in interpretation of breast MRI, a computer-aided diagnosis (CAD) system that can automatically analyze lesion features to differentiate between malignant and benign lesions would be very useful.

The general approach of breast CAD to predict a dichotomous outcome, malignant versus benign, involves applying computer algorithms for tumor characterization and then developing models using techniques including linear discriminate analysis (12), logistic regression analysis (13), and artificial neural networks (ANN) (14) for classifying a lesion as either malignant or benign. Because of inherent differences in these techniques, it is of interest to compare the diagnostic performance of analysis methods. In a review of 28 studies comparing ANN with other statistical approaches, Sargent found that ANN outperformed regression analysis in 36% of cases, ANN was outperformed by regression analysis in 14% of cases, and that similar results were obtained in 50% of cases (15). The advantages of ANN for predicting a dichotomous outcome include a requirement of less formal statistical training, an ability to implicitly detect complex nonlinear relationships between dependent and predictor variables, consideration of all possible interactions between predictor variables, and the availability of multiple training algorithms. Disadvantages include the “black box” nature of the ANN procedure, a greater computational burden, the tendency for overfitting, and the empirical nature of model development (16). Logistic regression is superior for examining possible causal relationships between independent and dependent variables, and understanding the effect of predictors on outcome variables (16). In summary, there is no universal approach to select models for data classification; evaluation of tasks on a case-by-case basis is necessary.

To our knowledge, only one study has compared ANN and logistic regression for breast cancer diagnosis using sonograms by Song et al. to differentiate between 24 malignant and 30 benign lesions (17). No difference was found in the diagnostic performance of ANN and logistic regression when measured by the area under the ROC curve. We have developed a CAD that incorporates a clustering-based algorithm for lesion segmentation and derives a full panel of quantitative morphological and texture descriptors for lesion characterization (18). In that study the ANN was used to select diagnostic features. In the present work we investigated and compared the diagnostic performance using the ANN and logistic regression.

Two aims were studied. Firstly we applied ANN to select features and investigated the diagnostic performance of different feature sets (models) selected based on the morphology and texture of the lesion. While ANN can select features robustly, the non-linear diagnostic model makes the weighting of each individual feature not interpretable. The logistic regression was applied to gain more insightful understanding for how each selected feature can be interpreted. Secondly, we illustrated the use of logistic regression for feature selection. The following statistical procedures were applied sequentially: 1) transformation of feature values to induce normality, 2) consideration of strategies for model selection and validation, 3) assessment of correlation between features to reduce collinearity, 4) reduction of the number of variables in models to avoid overfitting, and 5) evaluation and comparison of the diagnostic

performance of classifiers. The selected features were compared to those selected by ANN. Also the diagnostic performance achieved by each selected model was evaluated and compared to that achieved by ANN.

MATERIALS AND METHODS

Malignant and Benign Lesion Database

The database analyzed in this study included 43 malignant and 28 benign lesions, the same database as reported in a previous study (18). The 43 patients with malignant lesions were from 29 to 76 years old with mean (\pm SD) 48 ± 9 yr and median 48 yr. The 28 patients with benign lesions were from 21 to 74 yr, with mean 45 ± 7 yr. and median 45 yr. The MRI studies were performed on a Philips Eclipse 1.5T scanner (Cleveland, OH). The images were acquired using a T1-weighted 3D SPGR (RF-FAST) pulse sequence, with TR=8.1 ms, TE=4.0 ms, flip angle=20°, matrix size=256×128, FOV (Field of View) varying between 32 and 38 cm. The lesion was identified based on contrast enhanced images at 1-min after injection of the MR contrast medium Omniscan® (0.1 mmol/kg body weight).

The lesion was segmented on contrast-enhanced images. The detailed procedures have been reported previously (18). The operator inspected the lesion presentations on MRI, and determined the beginning and the ending image slices that contain the lesion, then one square box was placed on the lesion on one image slice to indicate the lesion location. With these inputs the 3-dimensional boundary of the lesion (or, lesion ROI- Region Of Interest) was obtained using an automated computer program based on the fuzzy c-means (FCM) clustering algorithm (1). After the 3-dimensional ROI for one lesion was defined, computer algorithms were used to extract quantitative features representing the morphological and texture properties of this lesion.

Eight morphological parameters were obtained for each tumor, including: volume, surface area, compactness, normalized radial length (NRL) mean, sphericity, NRL entropy, NRL ratio, and roughness, were obtained. Texture is a repeating pattern of local variations in image intensity, and is characterized by the spatial distribution of intensity levels in a neighborhood. Therefore, the texture features analyzed within the contrast enhanced lesion represent the distribution of enhancements. Ten GLCM texture features (energy, maximum probability, contrast, homogeneity, entropy, correlation, sum average, sum variance, difference average, and difference variance), as defined by Haralick and colleagues, were obtained for each lesion (19).

In addition, commonly used Laws' texture features were also obtained. The Laws' features were computed by first applying small convolution kernels to a digital image and then performing a nonlinear windowing operation. The 2D convolution kernels were generated from a set of 1D convolution kernels: L5=[1 4 6 4 1], E5=[-1 -2 0 2 1], S5=[-1 0 2 0 -1], R5=[1 -4 6 -4 1], W5=[-1 2 0 -2 -1]. These mnemonics stood for Level, Edge, Spot, Ripple and Wave. 2D convolution kernels such as Law_LE were obtained by convolving a vertical L5 kernel with a horizontal E5 kernel. In total, a set of 14 texture features that were rotationally invariant were obtained (Law_LE, Law_LS, Law_LW, Law_LR, Law_EE, Law_ES, Law_EW, Law_ER, Law_SS, Law_SW, Law_SR, Law_WW, Law_WR, and Law_RR) (20).

Artificial Neural Network (ANN) for diagnostic feature selection

A three-layer back-propagation neural network, known as multi-layer perceptrons (MLP) artificial neural network (ANN) was utilized to obtain optimal classifiers. The three-layer topology has an input layer, one hidden layer, and an output layer. The number of nodes in the input corresponds to the number of input variables. The output layer contains one node with

values from 0 to 1 indicating level of malignancy, where 0 means absolutely benign and 1 means absolutely malignant. The number of hidden nodes is usually determined by a number of trial-and-error runs. Different neural network architectures with hidden nodes from 2 to 31 were tested. A stochastic gradient descent with the mean squared error function was used as the learning algorithm. The optimal architecture was chosen as the one for which the validation error was the lowest. With the determined number of hidden nodes, both the learning rate (0.1–0.7) and the momentum coefficient (0.2–0.8) were varied during network training to ensure a high probability of global network convergence. For training, criteria for convergence was met with a root mean squared error less than or equal to 0.0001 or a maximum of 1000 iterations. With the determined number of hidden nodes, both the learning rate, (0.1–0.7), and the momentum coefficient, (0.2–0.8), were varied during network training to ensure a high probability of global network convergence.

After the topology was chosen, feature selection was performed to find potential predictors within each set of morphology, GLCM texture, and Laws' texture features. The features were selected using the LNKnet (<http://www.ll.mit.edu/IST/lknnet/>) package in order to identify those yielding maximum discrimination capability thus achieving the optimal diagnostic performance. Each feature from 71 lesions was transformed to have zero mean and unit variance before training; after determination of the mean and standard deviation (STD) for a particular feature with $n=71$, the transformation for the i^{th} value of a feature is $z_i = (\text{feature value} - \text{mean})/\text{STD}$. This transformation is known as a z-score.

A forward search strategy was applied to find the optimal feature subset, which was obtained when the trained classifier produced the least error rate. The k -fold stratified cross-validation technique with $k=4$ was applied to test the performance of the generated classifiers. Weights and bias of the neural network were determined by a two-phase training procedure. The first phase had 30 iterations of back propagation, and the second phase had a longer run of conjugated gradient descent to ensure full convergence. The logistic sigmoid function was used to interpret the output variation in terms of probability of class membership within the range (0 to 1). To control for overfitting, the potential feature set was limited to no more than 4 in each category which had the least error. After the features from each of three categories (morphology, GLCM, and Laws') were individually obtained (Models A, B, and C respectively), all selected predictors were considered in a combined model (Model D), and a final model, restricted to contain no more than four features (Model E), was selected using ANN. Specifically, these models consisted of Model A with three features selected from eight morphology features, Model B with three features selected from 10 GMLM features, and model C with one selected from 14 Laws' features, Model D consisting of the combined seven selected features, and Model E consisting of four features selected from the combined seven features.

Logistic Regression Model

For this study the binary response variable of interest is the presence of a malignant breast lesion, coded 1=yes and 0=no. Note that the mean of the response values for a sample of lesions is the same as the proportion of lesions that are malignant. A logistic regression model that predicts a transformation of the response variable is used, $\text{logit}(p)$. Let p = probability that a lesion is malignant and $1-p$ = probability that a lesion is benign. Then the log odds is defined as

$$\text{logit}(p) = \log_e \left(\frac{p}{1-p} \right) \quad (1)$$

The logistic regression model enables prediction of the probability of a malignant breast lesion in relation to m lesion features, $x_1, x_2, x_3, \dots, x_m$ from an equation of the form

$$\log_e \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_m x_m \quad (2)$$

where β_0 is the intercept term, and $\beta_1, \beta_2, \beta_3, \dots, \beta_m$ are the coefficients in the model associated with the m lesion features. It is assumed that the lesion features are linearly related to the log odds of the response. For a continuous lesion feature, the odds ratio for malignant versus benign lesions can be estimated as the exponentiation of the associated coefficient in the model (21).

Evaluation of ANN models using Logistic Regression

To address the first aim of the study, the selected diagnostic features in the five final models (A-E) by ANN were analyzed using logistic regression. Lesions were classified as benign or malignant on the basis of increasing threshold probabilities of malignancy. The ROC curve was constructed from the full range of probability thresholds with corresponding data points (sensitivity, 1-specificity) and the AUC was calculated. A nonparametric approach was used to compute the 95% confidence interval for the AUC (22,23). The classification of lesions on the basis of probability of being malignant > 0.5 was used to compare results for logistic regression with ANN for accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Logistic Regression Analysis for Diagnostic Feature Selection

To address the second aim of the study, logistic regression was used for initial selection of predictive models from feature z -scores. Descriptive statistics were calculated for features obtained from 71 lesions. The k -fold stratified cross-validation was applied with $k=4$. Forty-three malignant and twenty-eight benign lesions were separately and randomly assigned into four sub-cohorts so that each sub-cohort contained approximately the same proportion of malignant cases as the original cohort. To build predictive models, four sets of any three sub-cohorts were combined and used as a training cohort, with the remaining sub-cohort used as the corresponding validating cohort. These were labeled training cohorts 1–4 with corresponding validating cohorts 1–4.

The analysis method is now summarized. For each training set, logistic regression analysis was applied to examine the association between each individual feature and the dichotomous outcome, malignant lesion versus benign lesion. For example, univariate logistic regression models were formed, one for each of the eight morphological features. For each model, the likelihood ratio statistic was computed and the Bonferroni-Holm multiple comparisons method was applied to indicate statistically significant features at a 5% experimentwise significance level. These procedures were repeated for the 10 GLCM texture features and for the 14 Laws' texture features and the statistically significant features for each group were retained.

Next, for each training set, stepwise logistic model selection was applied to each set of features. For example, stepwise logistic regression was applied to the eight morphological features as predictors. The Score chi-square statistics were calculated and used to determine addition and removal of variables at the 0.05 significance level. The stepwise procedure was repeated for the 10 GLCM texture features and for the 14 Laws' texture features. Due to the small sample size of the cohorts, at most two predictors were retained from each set of features.

Pearson's correlation coefficient was computed for the pairs of predictors retained on the basis of univariate and stepwise modeling. Multivariate models were then built considering the combined morphology, GLCM texture and Laws' texture features that did not have statistically significant pairwise correlation. For each multivariate model, the variance inflation factor

(24) for features in a model were examined with the objective of minimizing multicollinearity. No effect of multicollinearity is indicated by a VIF of zero. A common rule of thumb is that if the variance inflation is >5 , then multicollinearity is high.

For each training set, a list of models was generated in ascending order of estimated AUC. The c index of concordance statistic was calculated to compare training datasets with regard to the receiver operating characteristic area (25,26) and the 95% confidence interval of the average of concordance statistics was obtained for each model. Moreover, the Hosmer-Lemeshow goodness-of-fit test was performed for each model (26,27). The models selected from each training cohort were applied to the corresponding validating sub-cohort.

Finally, the models with the greatest AUC for each validating sub-cohort were applied to the entire cohort. For each of these models, the accuracy, sensitivity, specificity, PPV, and NPV were calculated.

The odds of a lesion to be malignant was computed for every increase of 1 SD in a feature, adjusted for differences in the other features in the model. For the comparison, the same analysis method was used after Box-Cox transformations were applied to induce normality, as needed.

RESULTS

Artificial Neural Network Diagnostic Feature Selection and Evaluation

Before ANN training, features were transformed to z-scores based on the mean and standard deviation of the entire set of 71 lesions. A three layer ANN with 5 hidden nodes in the hidden layer was chosen after a number of trial-and-error runs. The diagnostic measures used to assess differentiation between 43 malignant and 28 benign lesions are summarized in Table 1. Considering the eight morphology features, the classifier selected by ANN included three parameters: lesion volume, NRL entropy, and compactness (Model A). Using these three features for ROC analysis of the entire cohort of 71 lesions, an AUC of 0.80 and accuracy of 0.77 was achieved. Considering the ten GLCM texture features, the selected parameters were: gray level energy, gray level sum average, and homogeneity (Model B), with an AUC of 0.81 and accuracy of 0.73. From fourteen Laws' energy features, the best classifier contained only one parameter, Law_LS (Model C), with an AUC of 0.70 in the ROC analysis and accuracy of 0.65. Applying ANN to the combined set of seven features, including compactness, lesion volume, and NRL entropy, energy, gray level sum average, homogeneity, and the Law_LS parameter (Model D), the resulting AUC and accuracy were improved to 0.87 and 0.79, respectively, with positive and negative predictive values of 0.83 and 0.72. When these seven features were considered by a further ANN selection process, the final four selected features were: compactness, energy, homogeneity, and Law_LS, (Model E) and achieved an AUC of 0.82, accuracy of 0.76, PPV of 0.78, and NPV of 0.72.

Logistic regression was applied to each of the models selected by ANN for the cohort of 71 lesions, classifying a lesion as malignant if the threshold probability of malignancy was >0.5 . The diagnostic measures are also displayed in Table 1 for comparison with the ANN results. The values for diagnostic measures are similar to those recorded from the ANN technique. For example, the 3-feature morphology model of volume, NRL entropy, and compactness (Table 1, model A) has an AUC of 0.80 recorded by the ANN technique, and the same AUC of 0.80 (95% CI, 0.686 to 0.912) computed using logistic regression. The 95% confidence interval for the AUC from logistic regression contains the 0.80 value recorded from ANN. Diagnostic accuracy of 0.77, recorded by ANN, was higher than the value of 0.76 computed from logistic regression. Similarly, the final 4-feature model of compactness, energy, homogeneity, and Law_LS, (Table 1, Model E), had an AUC of 0.82 recorded by the ANN technique, as compared to 0.80 (95% CI, 0.688 to 0.905) computed on the basis of logistic regression. The 95%

McLaren et al.

Page 7

confidence interval for AUC contains the computed value of 0.82 from ANN. Figure 1 displays the ROC curves for the 4-feature model (E) analyzed using the ANN and logistic regression. The diagnostic accuracy of 0.76 recorded by ANN was slightly higher than 0.72 computed by logistic regression.

For the outcome of presence of a malignant lesion and modeling of z-scores from feature predictors with $n=71$, the logistic regression equation was

$$\text{logit}(p) = 0.69 + 1.69 \text{ Compactness} - 0.61 \text{ Energy} + 0.24 \text{ Homogeneity} - 0.73 \text{ Law_LS} \quad (3)$$

Law_LS was the most predictive variable in the model (Wald statistic, $P=0.028$). The P-value of the Hosmer-Lemeshow test was 0.76 indicating a reasonably good fit of the model to the data. However, the adjusted R^2 for the model was 0.30 which indicates that only 30% of the total variation in values could be explained by the predictor features in the model (28). The estimated odds ratio for Law_LS was calculated as the exponentiation of the associated estimated coefficient, $\exp(-0.73) = 0.48$. Thus the odds of a malignant lesion decreased by 48% (95% CI, 25% to 92%) for every increase of 1 SD in the Law_LS feature, adjusted for differences in compactness, energy, and homogeneity. For example, comparing a lesion with Law_LS of 0.19, one standard deviation above the mean of 0.13 for the set of 71 lesions, to a lesion Law_LS of 0.25, two standard deviations above the mean, the odds of the second lesion being malignant are about one-half of that of the first lesion, assuming that they have similar values for compactness, energy, and homogeneity.

Diagnostic Feature Selection using Logistic Regression Analysis

To address the second aim of the study, we applied logistic regression to z-scores for morphology, GLCM texture, and Laws' texture features and followed the modeling strategy described in Materials and Methods. Within training sets, features preselected for consideration by multivariate models had no statistically significant correlation between pairs of features ($P>0.13$ for all). Table 2 displays the best single model selected from each training cohort. For each training cohort, the model chosen had the highest AUC from among those applied to the corresponding validation cohort. For a particular model, the diagnostic values presented in the table were calculated when the model was used to reclassify the full data set of 71 lesions on the basis of probability of being benign or malignant. Each of the four models (Table 2, training cohort 1–4) contained two morphological parameters, compactness and NRL Entropy. In addition, models chosen from training cohort 1 and 2 contained a GLCM texture feature. The model selected from training cohort 1 included energy; while the model from training cohort 2 included gray level sum average. Identical models were chosen from training cohorts 3 and 4 and contained the Laws' texture feature, Law_LW. When the chosen models were applied to the full data set of 71 lesions, all models had P-values for the likelihood ratio test less than 0.004 indicating the joint significance of all predictor features with respect to the outcome. Values for AUC varied from 0.75 for the model from training cohort 1 (compactness, NRL entropy, energy) to 0.80 for the model selected from training cohorts 3 and 4 (compactness, NRL entropy, Law_LW). However, the 95% confidence intervals for the AUC overlap, indicating no significant difference between AUC values. Estimated accuracy of models varied from 72% to 75%. The model chosen from training cohort 2 had the highest overall accuracy, correctly classifying 75% of lesions, and an AUC of 0.77 (95% CI, 0.660 to 0.880). This model was comprised of three features including compactness, NRL entropy, and gray level sum average. The model had high sensitivity, correctly identifying 91% of malignant lesions and moderate specificity, correctly identifying 50% of benign lesions. While the model selected from training cohort 1 (compactness, NRL entropy, and energy) had the highest sensitivity and correctly identified 95% of malignant cases, the specificity was only 39%.

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Coding Density

McLaren et al.

Page 8

Figure 2 displays the ROC curve, illustrated with a solid line, for the 3-feature model selected from training cohort 2 with the highest overall accuracy among validating cohorts, correctly classifying 75% of the 71 lesions. The model was comprised of two morphological parameters, compactness and NRL entropy, and one GLCM feature, gray level sum average (Table 2, Figure 2). Applied to the entire cohort of 71 lesions, there was no statistically significant correlation between pairs of variables ($P > 0.10$ for all). A good fit of the model to the data was demonstrated (Hosmer and Lemeshow goodness-of-fit statistic, $P = 0.88$) and a high concordance statistic (0.78; 95% CI, 0.755 to 0.813) implied a good ability to predict the lesion type. As expected, the average variance inflation factor (VIF) for the three features in the model was 1.02 indicating low effect of multicollinearity on the variance of model coefficients. The strongest predictor in the model was gray level sum average. The odds of a malignant lesion increased by 2.0 (95% CI, 1.14 to 3.68) for every increase of 1 SD in the texture feature, gray level sum average, adjusted for differences in morphology features, compactness and NRL entropy. The logistic regression equation was

$$\text{logit}(p) = 0.66 + 1.28 \text{ Compactness} - 0.66 \text{ NRL Entropy} + 0.72 \text{ Gray Level Sum Average} \quad (4)$$

Gray level sum average was the most predictive variable in the model (Wald statistic, $P = 0.016$). For this variable, the estimated odd ratio was $\exp(0.72) = 2.1$ (95% CI, 1.14 to 3.68), thus the odds of a malignant lesion increased by 2.1 for every increase of 1 SD in the gray level sum average feature, adjusted for differences in compactness and NRL entropy. For example, comparing a lesion with gray level sum average of 27.6, the mean for the 71 lesions, to a lesion with feature value of 34.2, one standard deviation above the mean, the odds of the second lesion being malignant are over twice that of the first lesion, assuming that they have similar values for compactness and NRL entropy.

We compared results of modeling the feature z -scores to those obtained after applying the Box-Cox method of selecting the most appropriate transformation to induce normality. A two-feature model of \ln compactness and Box-Cox transformed Law_LS had the highest accuracy of 0.79, with estimated AUC of 0.79 (95% CI, 0.672 to 0.898), sensitivity 0.88 and specificity 0.64. Figure 2 displays the ROC curves for the 2-feature model. The likelihood ratio chi-square for the model was statistically significant ($P < 0.0001$). As expected, the correlation between \ln compactness and transformed Law_LS was not statistically significant ($r = 0.196$, $P = 0.10$) and a good fit of the model to the data was demonstrated (Hosmer and Lemeshow Goodness-of-Fit, $P = 0.66$) with a concordance statistic 0.79 (95% C.I., 0.724 to 0.855). The average variance inflation factor was 1.04, indicating a low effect of multicollinearity on the variance of model coefficients. In the model, both features added information to the model (Wald statistic, $P < 0.007$ for each). For the outcome of presence of a malignant lesion, the logistic regression equation was

$$\text{logit}(p) = -5.650 + 0.666 (\ln \text{ Compactness}) - 266.50 ((\text{Law_LS}^{(0.1)}) - 1) / (1 / (0.1)) \quad (5)$$

For \ln compactness, the estimated odds ratio was $\exp(0.666) = 1.9$. This indicates that the odds of a malignant lesion increased by 1.9 (95% CI, 1.20 to 3.56) for every increase of 1 unit in the natural log of the morphology feature, compactness, adjusted for differences in the Laws' feature, Law_LS. For example, comparing a lesion with \ln compactness of 3.0, one standard deviation above the mean of 1.6 for the set of 71 lesions, to that of a lesion with \ln compactness of 4.5, two standard deviations above the mean, the second lesion would be nearly twice as likely to be malignant as first lesion, assuming equal values for Law_LS.

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

DISCUSSION

The primary aim of the study was to compare ANN and logistic regression analysis for lesion classification to differentiate between malignant and benign breast lesions in patients. Using our dataset of 71 lesions, the ANN procedure was applied to select the best classifiers for morphology and texture (GLCM and Laws') category features. The three selected morphology features (volume, NRL entropy, compactness) achieved a moderate AUC of 0.80 and estimated accuracy of 0.77. The three selected GLCM features (energy, gray level sum average, and homogeneity) achieved higher AUC (0.81) and estimated accuracy (0.73). Only one Laws' feature (Law_LS) was selected, and achieved lower AUC (0.70) and lower accuracy (0.65). When all seven features were combined the model achieved an improved AUC of 0.87 and estimated accuracy of 0.79. Submitting these seven features into another ANN selection, resulted in selection of four features, one morphology feature (compactness), two GLCM features (energy, homogeneity), and one Laws' texture feature (Law_LS) with an AUC of 0.82 and accuracy of 0.76. These results demonstrate that it is possible to use ANN to select the best combined indicators to predict tumor malignancy.

For the 4-feature model (Table 1, Model E), logistic regression analysis revealed that Law_LS was the most predictive variable in the model (Wald statistic, $P=0.028$). The estimated odds ratio for Law_LS was calculated as $\exp(-0.73) = 0.48$. Thus the odds of a malignant lesion decreased by 48% (95% CI, 25% to 92%) for every increase of 1 SD in the Law_LS feature, adjusted for differences in compactness, energy, and homogeneity. This result provides an illustration of the complementary manner in which ANN and logistic regression can be used (15). While ANN is more robust in that features are selected with minimal intellectual judgment of the operator, logistic regression can provide more insight and understanding into the relationship between selected features and the outcome.

Artificial neural networks have been used elsewhere in clinical data modeling, and similar results with that of regression modeling techniques were demonstrated (29–33). Compared to logistic regression modeling, ANN was found to have higher prediction rates in complex and non-linear relationships among a large number of variables; however even when the difference was significant (due to very large sample size) the improved performance was only marginal. Nilsson et al. used data from 18,362 patients undergoing cardiac surgery to predict the operative mortality. ANN selected 34 of the total 72 risk variables as relevant for mortality prediction. The area under ROC curve for ANN (0.81) was larger than that of the logistic regression model (0.79, $P=0.0001$) with the same 34 top-ranked risk variables (29). Delen and colleagues acquired a large dataset (202,932 cases with 17 variables) to predict five-year breast cancer survival using 10-fold cross-validation. The results indicated that when all variables were used, the ANN estimated area under the ROC curve with 0.91, compared to 0.89 computed by logistic regression (31). Lundin et al. tried to predict five-year breast cancer survival using data from 951 breast cancer patients. Using eight input variables, the ANN and logistic regression models achieved similar values for AUC of 0.901 and 0.897 respectively (33). Jaimes et al. (32) and Clermont et al. (30) also found that both ANN and logistic regression have similar performance when considering a small number of variables.

Advantages of neural network analysis are that few prior assumptions or knowledge about data distributions are required, so knowledge about complex variable transformations is not needed before training, and the search for the optimal diagnostic classifier involves minimal user input. Another advantage is that ANN has the capacity to model complex nonlinear relationships between independent and predictor variables, allowing the inclusion of a large number of variables. A disadvantage of ANN is the long training process and requirement of an experienced operator to determine the optimal network topology. The major factor that needs to be experimentally determined is the number of hidden layer nodes. If too few hidden nodes

are used, proper training is impeded. If too many are used, the neural network is over-trained. In our study the number of hidden nodes was determined by a number of trial-and-error runs. Another limitation of the ANN technique is the poor interpretability of selected models. Neither standardized coefficients nor odd ratios corresponding to selected variable can be calculated and presented as in regression models. Logistic regression can be applied in a complementary manner to provide this information, thus overcoming the problem. Furthermore, the technique can be used for hypothesis testing regarding univariate and multivariate associations between predictor variables and the outcome of interest (15) and to enhance understanding and interpretation of the effect of predictor variables on the response (16). The logistic regression analysis may be preferred to ANN due to improved interpretation of individual predictors.

Both ANN and logistic regression are subject to issues of overfitting, assessing model convergence, and collinearity that affect the generalization of results (15). Methods to avoid overfitting include cross-validation as applied in our study (34). Cross-validation serves to check internal validity (reproducibility) (35). Leave-one-out cross validation, although almost unbiased, may have high variance leading to unreliable estimates (36). Kohavi studied cross-validation for accuracy of estimation and model selection and recommends *k*-fold stratified cross validation for model selection on the basis of stability of predictions and accuracy, when compared to leave-one-out cross validation (34,37). As recommended, in this study we use 4-fold cross validation for ANN and logistic regression modeling. Finally, it is important to note that ANN does not strictly check collinearity among features during the selection process and collinearity can affect the variance of model estimates. Since the potential for collinearity among the features is not specifically taken into account, this can lead to stability problems (38–40).

The second aim of the study was to illustrate the use of logistic regression for feature selection. We examined differences between methods of standardizing features and addressed the issue of potential collinearity by incorporating statistical testing for correlation between variables to pre-select variables for further modeling. As illustrated in Table 2, the logistic regression model of feature z-scores with the highest estimated accuracy of 0.75 and AUC (0.77; 95% CI, 0.660 to 0.880) included compactness, NRL entropy, and gray level sum average (Figure 2). The model of Box-Cox transformed values with the highest accuracy of 0.79 contained two features, ln compactness and Box-Cox transformed Law_LS, with estimated AUC of 0.79 (95% CI, 0.672 to 0.898), sensitivity 0.88 and specificity 0.64 (Figure 2). The results suggest that the diagnostic performance of the models selected by logistic regression was comparable to that of ANN; also that the method of standardization may improve on model selection.

Regarding the generalizability of our results, although we analyzed a relatively small dataset, the selected features have been well accepted and commonly used in the development of breast CAD systems. Using this same dataset, we previously attempted to establish the association between the extracted quantitative features and the lesion phenotype appearance on MRI as described in the BI-RADS breast MRI lexicon (18). For example, the compactness morphological parameter is strongly linked to the shape and margin of the lesion, and the GLCM texture features are associated with the degree of the enhancement and the heterogeneous enhancement patterns within the lesion. These BI-RADS descriptors are well-established diagnostic features. Therefore, although the generalization of the selected quantitative features need validation using independent datasets, they are generalizable in the sense that they are closely related to visual diagnostic features commonly used by radiologists.

In summary, we have shown that the diagnostic performance of models selected by ANN and logistic regression was similar and the analytic methods were found to be roughly equivalent in terms of predictive ability when a small number of variables were chosen. We have emphasized interpretation of the predictors in the model and illustrated comparison of lesion

features in terms of the odds of a lesion being malignant, enhancing the usefulness of the logistic regression modeling. The ANN methodology is more robust (i.e. it does not require a high level of operator judgment), and it utilizes a sophisticated non-linear model to achieve a high diagnostic performance. On the other hand, logistic regression may generate many sets of models that yield similar diagnostic performance, and the operator will need to make intellectual judgments to select the best model(s). The modeling strategy provided in the present work requires statistical judgment and thus may be more difficult to implement in a large dataset that has a many variables compared to the “black box” ANN approach. Nonetheless, logistic regression analysis provides insightful information to enhance interpretation of the model features. Finally, many diagnostic models (feature sets) could be selected using ANN and logistic regression based on cross-validation within one dataset; and the ultimate diagnostic value of these models will have to be determined in an independent validation dataset.

Acknowledgments

This work was supported in part by NIH/NCI R01 CA90437 (O. Nalcioglu), CA121568 (M-Y Su), the California Breast Cancer Program grant #9WB-002 (M-Y Su), and the UC Irvine Cancer Center Support Grant No. 2P30CA062203-13S (F.L. Meyskens, Jr.)

REFERENCES

1. Chen W, Giger ML, Bick U. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Acad Radiol* 2006;13:63–72. [PubMed: 16399033]
2. Chen W, Giger ML, Bick U, et al. Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI. *Med Phys* 2006;33:2878–2887. [PubMed: 16964864]
3. Chen W, Giger ML, Lan L, et al. Computerized interpretation of breast MRI: investigation of enhancement-variance dynamics. *Med Phys* 2004;31:1076–1082. [PubMed: 15191295]
4. Liney GP, Sreenivas M, Gibbs P, et al. Breast lesion analysis of shape technique: semiautomated vs. manual morphological description. *J Magn Reson Imaging* 2006;23:493–498. [PubMed: 16523479]
5. Meinel LA, Stolpen AH, Berbaum KS, et al. Breast MRI lesion classification: improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system. *J Magn Reson Imaging* 2007;25:89–95. [PubMed: 17154399]
6. Esserman L, Hylton N, Yassa L, et al. Utility of magnetic resonance imaging in the management of breast cancer: evidence for improved preoperative staging. *J Clin Oncol* 1999;17:110–119. [PubMed: 10458224]
7. Fischer U, Kopka L, Grabbe E. Breast carcinoma: effect of preoperative contrast-enhanced MR imaging on the therapeutic approach. *Radiology* 1999;213:881–888. [PubMed: 10580970]
8. Mumtaz H, Hall-Craggs MA, Davidson T, et al. Staging of symptomatic primary breast cancer with MR imaging. *AJR Am J Roentgenol* 1997;169:417–424. [PubMed: 9242745]
9. Rieber A, Schirmer H, Gabelmann A, et al. Pre-operative staging of invasive breast cancer with MR mammography and/or PET: boon or bunk? *Br J Radiol* 2002;75:789–798. [PubMed: 12381687]
10. Schelfout K, Van Goethem M, Kersschot E, et al. Contrast-enhanced MR imaging of breast lesions and effect on treatment. *Eur J Surg Oncol* 2004;30:501–507. [PubMed: 15135477]
11. Zhang Y, Fukatsu H, Naganawa S, et al. The role of contrast-enhanced MR mammography for determining candidates for breast conservation surgery. *Breast Cancer* 2002;9:231–239. [PubMed: 12185335]
12. Gilhuijs KG, Giger ML, Bick U. Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. *Med Phys* 1998;25:1647–1654. [PubMed: 9775369]
13. Chou YH, Tiu CM, Hung GS, et al. Stepwise logistic regression analysis of tumor contour features for breast ultrasound diagnosis. *Ultrasound Med Biol* 2001;27:1493–1498. [PubMed: 11750748]
14. Chan HP, Sahiner B, Petrick N, et al. Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network. *Phys Med Biol* 1997;42:549–567. [PubMed: 9080535]

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

15. Sargent DJ. Comparison of artificial neural networks with other statistical approaches. *Cancer* 2001;91:1636–1642. [PubMed: 11309761]
16. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996;49:1225–1231. [PubMed: 8892489]
17. Song JH, Venkatesh SS, Conant EA, et al. Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. *Acad Radiol* 2005;12:487–495. [PubMed: 15831423]
18. Nie K, Chen J-H, Yu HJ, et al. Quantitative Analysis of Lesion Morphology and Texture Features for Diagnostic Prediction in Breast MRI. *Acad Radiol*. 2008 (in press).
19. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern* 1973;SMC-3:610–621.
20. Laws, KI. Society of Photo-Optical Instrumentation Engineers. San Diego, CA: Image processing for missile guidance; 1980. Rapid texture identification; p. 376-380.
21. Altman, DG. London: Chapman & Hall; 1991. Practical statistics for medical research.
22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–845. [PubMed: 3203132]
23. Puri, ML.; Sen, PK. New York: Wiley; 1971. Nonparametric Methods in Multivariate Analysis.
24. Neter, J.; Kutner, M.; Nachtsheim, C., et al. New York: WCB McGraw-Hill; 1996. Applied linear statistical models.
25. Harrell FE, Cadiff RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA* 1982;247:2543–2546. [PubMed: 7069920]
26. Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer. *J Natl Cancer Inst* 2007;99:715–726. [PubMed: 17470739]
27. Hosmer, DW.; Lemeshow, S. New York: John Wiley & Sons; 1989. Applied Logistic Regression.
28. Nagelkerke NJD. A Note on a General Definition of the Coefficient of Determination. *Biometrika* 1991;78:691–692.
29. Nilsson J, Ohlsson M, Thulin L, et al. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *J Thorac Cardiovasc Surg* 2006;132:12–19. [PubMed: 16798296]
30. Clermont G, Angus DC, DiRusso SM, et al. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit Care Med* 2001;29:291–296. [PubMed: 11246308]
31. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med* 2005;34:113–127. [PubMed: 15894176]
32. Jaimes F, Farbiarz J, Alvarez D, et al. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Crit Care* 2005;9:R150–R156. [PubMed: 15774048]
33. Lundin M, Lundin J, Burke HB, et al. Artificial neural networks applied to survival prediction in breast cancer. *Oncology* 1999;57:281–286. [PubMed: 10575312]
34. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. The Fourteenth International Joint Conference on Artificial Intelligence Morgan Kaufmann; San Mateo. 1995. p. 1137-1143.
35. Terrin N, Schmid CH, Griffith JL, et al. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *J Clin Epidemiol* 2003;56:721–729. [PubMed: 12954463]
36. Efron B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J Am Stat Assoc* 1983;78:316–331.
37. Breiman L, Spector P. Submodel Selection and Evaluation in Regression. The X-Random Case. *Int Stat Rev* 1992;60:291–319.
38. Martens, H.; Næs, T. Chichester: Wiley; 1989. Multivariate Calibration.
39. Næs T, Mevik B-H. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics* 2001;15:413–426.

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

McLaren et al.

Page 13

40. Weisberg, S. New York: Wiley; 1985. Applied Linear Regression.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Colling Density

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

McLaren et al.

Page 14

Coding: Density

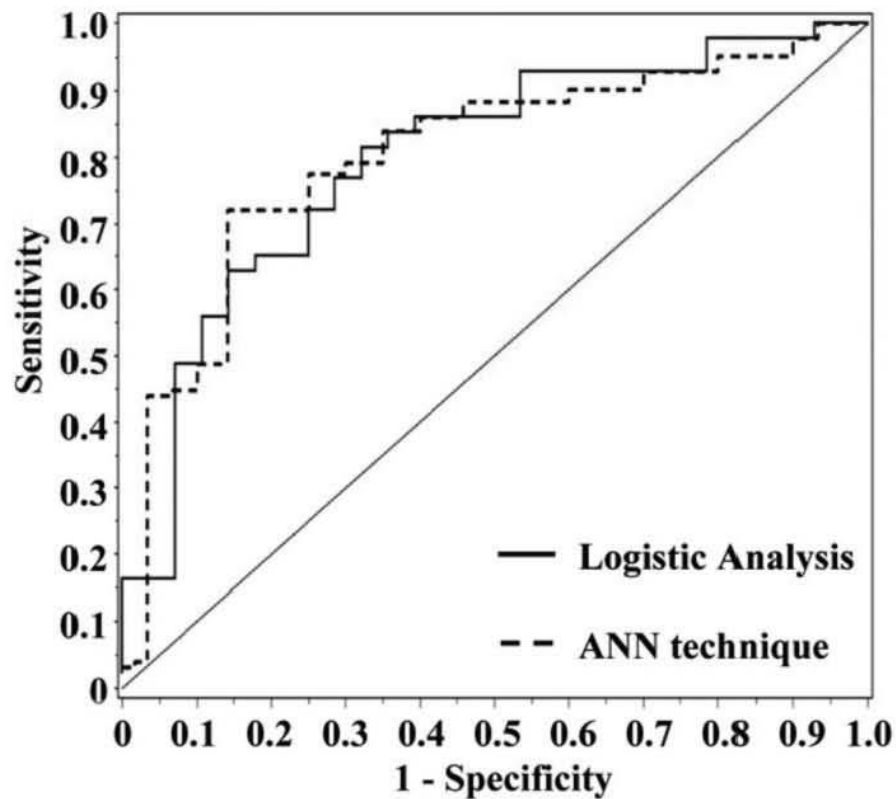


Figure 1. The dashed line represents the ROC curve for ANN modeling of z-scores (Table 1, Model E; Compactness, Energy, Homogeneity and Law_LS; AUC 0.82). The solid line represents the ROC curve for the four features as assessed by logistic regression (Table 1, Model E, AUC 0.80).

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

McLaren et al.

Page 15

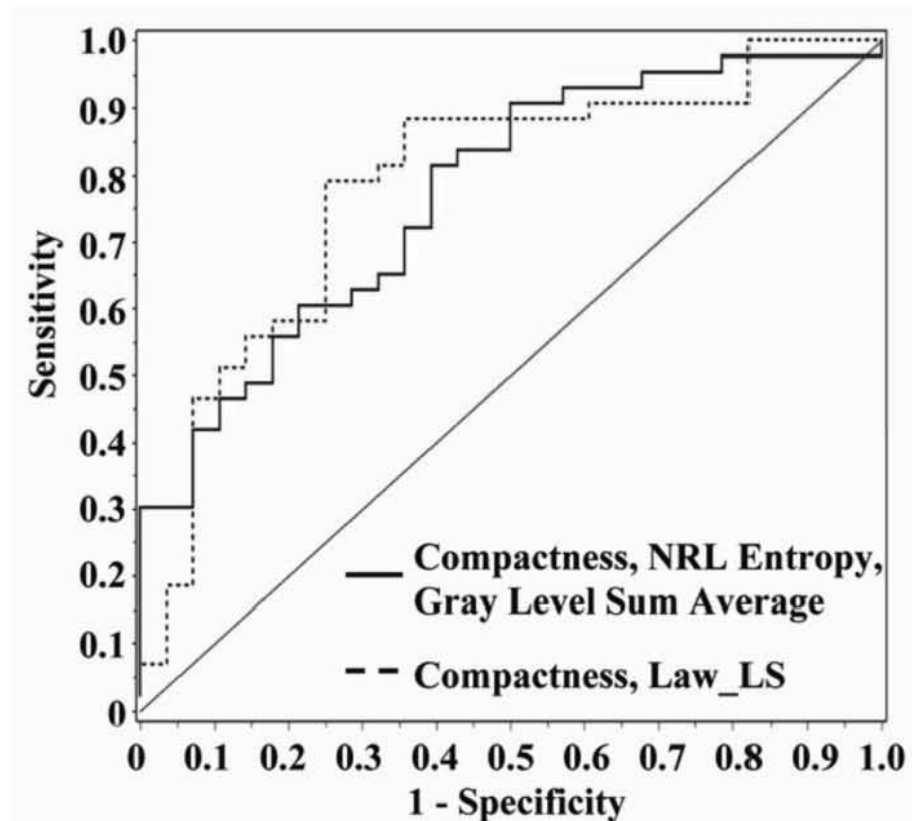


Figure 2.

The solid line represents the ROC curve for logistic regression modeling of z-scores for compactness, NRL entropy and gray level sum average (AUC 0.77; 95% CI, 0.660 to 0.880). The dashed line represents logistic regression model of Box-Cox transformed values for compactness and Law_LS (AUC 0.79; 95% CI, 0.672 to 0.898).

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Coding Density

McLaren et al.

Table 1
Diagnostic evaluation of models selected using the artificial neural network (ANN) technique. For each model the corresponding logistic regression equation also was applied to data from the full cohort (N = 71; Malignant = 43; Benign = 28).

Model	Imaging Descriptors ^a	Method	Accuracy ¹ (%)	Sensitivity ² (%)	Specificity ³ (%)	PPV ⁴	NPV ⁵	95% CI	
								Estimated Area under ROC	AUC
A	Morphology (3 selected from 8) Volume, NRL Entropy, Compactness	ANN	77	88	61	0.78	0.77	0.80	
		Logistic Regression	76	93	50	0.74	0.82	0.80	0.686
B	GLCM (3 selected from 10) Energy, Gray Level Sum Average, Homogeneity	ANN	73	84	57	0.75	0.70	0.81	
		Logistic Regression	68	86	39	0.69	0.65	0.77	0.660
C	LAAs (only 1 selected from 14) Law_LLS	ANN	65	84	36	0.67	0.59	0.70	
		Logistic Regression	65	86	32	0.66	0.60	0.70	0.573
D	Combining all 7 selected features Volume, NRL Entropy, Compactness, Energy, Gray Level Sum Average, Homogeneity, Law_LLS	ANN	79	81	75	0.83	0.72	0.87	
		Logistic Regression	80	86	71	0.82	0.77	0.86	0.772
E	Final (4 selected from 7) Compactness, Energy, Homogeneity, Law_LLS	ANN	76	84	64	0.78	0.72	0.82	
		Logistic Regression	72	86	50	0.73	0.70	0.80	0.688

¹Accuracy = (number of correctly identified cases / 71) × 100%

²Sensitivity = (number of correctly identified as malignant / 43) × 100%

³Specificity = (number of correctly identified as benign / 28) × 100%

⁴Positive Predictive Value (PPV) = number of lesions correctly identified as malignant / number of lesions identified as malignant

⁵Negative Predictive Value (NPV) = number of lesions correctly identified as benign / number of lesions identified as benign.

^aEach variable in the model was standardized by subtracting the mean and dividing by the standard deviation of data from 71 subjects.

Page 16

Coding Density

McLaren et al.

Table 2

Diagnostic evaluation of models selected using logistic regression of feature z-scores with 4-fold cross validation and applied to data from the full cohort (N = 71; Malignant = 43; Benign = 28).

Training Cohort	Model with highest AUC for corresponding validation cohort ^d	Likelihood ratio β^2	P-value	Accuracy ^f (%)	Sensitivity ² (%)	Specificity ³ (%)	PPV ⁴	NPV ⁵	95% CI	
									AUC	Estimated Area under ROC ⁶
1	Compactness, NRL Entropy, Energy	13.54	0.0036	73	95	39	0.71	0.85	0.75	0.626 - 0.879
2	Compactness, NRL Entropy, Gray Level Sum Average	17.86	0.0005	75	91	50	0.74	0.78	0.77	0.660 - 0.880
3,4	Compactness, NRL Entropy, Law_Law	19.16	0.0003	72	86	50	0.73	0.70	0.80	0.690 - 0.908

¹ Accuracy = (number of correctly identified cases / 71) × 100%.

² Sensitivity = (number of correctly identified as malignant / 43) × 100%.

³ Specificity = (number of correctly identified as benign / 28) × 100%.

⁴ Positive Predictive Value (PPV) = number of lesions correctly identified as malignant / number of lesions identified as malignant.

⁵ Negative Predictive Value (NPV) = number of lesions correctly identified as benign / number of lesions identified as benign.

⁶ Classification criteria: If the predicted value > 0.5, then the case was classified as malignant case by the model. If the predicted value < 0.5, then the case was classified as benign case by the model.

^d Each variable in the model was standardized by subtracting the mean and dividing by standard deviation of values from 71 subjects.

Acad Radiol. Author manuscript; available in PMC 2010 July 1.

Page 17

Coding Density

Appendix K

Clinical Applications Tally

Application Counts Study Application	Success		Grand Total
	Partial	Fully	
Breast Cancer	1	10	11
Metastatic Cancer	1	2	3
Epilepsy		3	3
Lung Cancer	1	1	2
Brain Cancer		2	2
Melanoma	1	1	2
Swallowing disorder		2	2
Liver Cancer		2	2
Thyroid Disease		2	2
Obstructive Sleep Apnea		1	1
Laryngopharyngeal Reflux		1	1
Risk of Death		1	1
Diabetes Mellitus		1	1
Chronic Subdural Hematoma	1		1
Diabetic Retinopathy		1	1
Pelvic Organ Prolapse		1	1
Disphonia	1		1
Interstitial Lung Disease		1	1
Endoscopic Therapy		1	1
Autoimmune Hemolytic Anemia		1	1
Basal Cell Carcinoma		1	1
Neuropathic Pain	1		1
Esophageal Cancer		1	1
Oral Cancer		1	1
Exertional Heat Illness		1	1
Post-op Nausea		1	1
Facial Pain	1		1
Skeletal Metastasis of PC		1	1
Febrile Neutropenia	1		1
Kidney Graft	1		1
Fibromyalgia Syndrome		1	1
Limb Fracture - Open		1	1
Focal Liver Disease		1	1
Chest Auscultation		1	1
Gestational Heart Rate		1	1

Application Counts (continued)		Success		Grand Total
Study Application	Partial	Fully		
Colorectal Cancer		1		1
Glaucoma		1		1
Obsessive-Compulsive Disorder		1		1
Headaches		1		1
Optic Nerve Disease		1		1
Heart Murmur		1		1
Parkinson's Disease		1		1
ADHD		1		1
Pneumoconiosis		1		1
Transmural Ischemia		1		1
Psychosocial Risk		1		1
Hemodialysis		1		1
Scoliosis		1		1
Hepatic Encephalopathy	1			1
Dengue Fever Risk		1		1
Hypogonadism	1			1
Heart Valve Disease		1		1
Acute Liver Failure		1		1
Hemiparetic Stroke		1		1
Grand Total	12	62		74
	16%	84%		

Appendix L

Actual Predictive Power by Study

Study ID	Performance Measure	Performance Value	Converted	Classification Successful?
3	Prediction %	98.10%	98.10%	Fully
8	Prediction %	91.60%	91.60%	Fully
9	Prediction %	88.38%	88.38%	Fully
12	AUROC	94.90%	94.90%	Fully
14	AUROC	94.16%	94.16%	Fully
16	AUROC	97.27%	97.27%	Fully
17	AUROC	88.00%	88.00%	Fully
19	Prediction %	98.03%	98.03%	Fully
20	Prediction %	99.12%	99.12%	Fully
21	Sens-Spec	98.00%/90.50%	94.25%	Fully
31	AUROC	98.00%	98.00%	Fully
33	Prediction %	95.50%	95.50%	Fully
38	Prediction %	99.67%	99.67%	Fully
43	Sens-Spec	100.00%/80.80%	90.40%	Fully
44	AUROC	99.00%	99.00%	Fully
47	Prediction %	64.00%	64.00%	Partial
49	AUROC	91.00%	91.00%	Fully
52	Prediction %	96.77%	96.77%	Fully
53	Prediction %	92.30%	92.30%	Fully
55	AUROC	72.50%	72.50%	Partial
77	Prediction %	96.86%	96.86%	Fully
99	AUROC	98.10%	98.10%	Fully
100	AUROC	83.35%	83.35%	Fully
103	Prediction %	96.79%	96.79%	Fully
112	Prediction %	90.00%	90.00%	Fully
121	AUROC	96.00%	96.00%	Fully
133	AUROC	95.00%	95.00%	Fully
161	Prediction %	92.86%	92.86%	Fully
164	Prediction %	96.00%	96.00%	Fully
165	Prediction %	93.00%	93.00%	Fully
166	Prediction %	83.10%	83.10%	Fully
169	AUROC	86.00%	86.00%	Fully
171	AUROC	95.00%	95.00%	Fully
172	Sens-Spec	75.00%/92.50%	83.75%	Partial
174	Sens-Spec	88.00%/83.00%	85.50%	Fully
178	Sens-Spec	99.20%/99.40%	99.30%	Fully
175	AUROC	85.60%	85.60%	Fully

Study ID	Performance Measure	Performance Value	Converted	Classification Successful?
<i>(continued)</i>				
181	Prediction %	98.30%	98.30%	Fully
182	Spearman	84.00%	84.00%	Fully
184	Prediction %	89.75%	89.75%	Fully
185	Prediction %	69.00%	69.00%	Partial
187	Prediction %	95.00%	95.00%	Fully
190	Prediction %	83.30%	83.30%	Fully
191	Prediction %	80.00%	80.00%	Fully
197	Sens-Spec	88.00%/80.00%	84.00%	Fully
204	Prediction %	83.00%	83.00%	Fully
207	Sens-Spec	90.00%/83.00%	86.50%	Fully
215	AUROC	87.00%	87.00%	Fully
216	AUROC	94.50%	94.50%	Fully
218	AUROC	76.70%	76.70%	Partial
230	Prediction %	95.00%	95.00%	Fully
233	AUROC	73.70%	73.70%	Partial
234	Prediction %	93.30%	93.30%	Fully
241	AUROC	88.40%	88.40%	Fully
245	AUROC	88.80%	88.80%	Fully
247	Prediction %	99.40%	99.40%	Fully
248	AUROC	78.10%	78.10%	Partial
250	AUROC	71.00%	71.00%	Partial
255	Prediction %	93.75%	93.75%	Fully
264	Prediction %	78.45%	78.45%	Partial
268	Prediction %	87.12%	87.12%	Fully
273	Prediction %	90.40%	90.40%	Fully
288	AUROC	90.90%	90.90%	Fully
289	AUROC	89.10%	89.10%	Fully
296	Prediction %	85.23%	85.23%	Fully
299	AUROC	95.64%	95.64%	Fully
319	Prediction %	91.11%	91.11%	Fully
323	AUROC	80.00%	80.00%	Fully
326	AUROC	88.70%	88.70%	Fully
329	Prediction %	99.00%	99.00%	Fully
337	Prediction %	73.05%	73.05%	Partial
344	AUROC	88.40%	88.40%	Fully
353	Sens-Spec	95.09%/92.19%	93.64%	Fully
361	Prediction %	88.00%	88.00%	Fully

Mean Predictive Power: 89.33%