A DIAGNOSTIC EXPERT SYSTEM, Iliad,
AS A QUALITY REVIEW SCREEN

by

Lee Min Lau

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Medical Informatics
The University of Utah

December 1994
THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a dissertation submitted by

Lee Min Lau

This dissertation has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

Chair: Homer R. Warner
A. Lloyd Poulsen
T. Allan Pryor
John W. Williamson

11/22/94
To the Graduate Council of the University of Utah:

I have read the dissertation of Lee Min Lau in its final form and have found that (1) its format, citations and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the supervisory committee and is ready for submission to The Graduate School.

Date

Homer R. Warner
Chair, Supervisory Committee

Approved for the Major Department

Homer R. Warner
Chair/Dean

Approved for the Graduate Council

Ann W. Hart
Dean of The Graduate School
ABSTRACT

A diagnostic expert system, Iliad, was evaluated as a quality review tool in the Medicare inpatient review performed by Peer Review Organizations (PROs).

The current PRO review appeared to underdetect diagnostic errors. The PRO review started with a nurse review using generic quality screens. Only cases flagged by the nurse review as containing a potential quality problem were then referred for a physician review. The quality screens focused only on treatment, discharge and documentation. Physician review, directed by the concerns raised in the nurse review, might also have failed to detect diagnostic errors. Thus, the expert diagnostic capability of Iliad could be useful as a nurse review screen for diagnostic errors. A discrepancy between Iliad’s diagnosis and that made by the attending physician for a case would flag a potential diagnostic error. The case could then be referred to a physician for confirmation.

Two experiments were performed. The first compared Iliad review with the current PRO review. On 100 Medicare inpatient cases flagged by the Utah Peer Review Organization (UPRO) nurse review as containing potential problems, and which contained diagnoses recognized by Iliad, Iliad review found 17% of the cases to have contained a diagnostic error, significantly higher than the 0% Diagnostic Error Rate (DER) found by UPRO. The Iliad nurse review True Positive Rate (TPR) was also significantly higher than that of the UPRO nurse review.

The second experiment compared Iliad review to the computerized UCDS (Unified Clinical Data Set) review undergoing testing at UPRO. In a random sample of 326 Utah Medicare inpatient cases containing diagnoses recognized by Iliad, Iliad review found a significantly higher DER of 6%, compared to that found by the UCDS review (0.8%). The Iliad nurse review TPR was also significantly higher than that of the UCDS nurse review.
The effects of varying the "strictness" of Iliad review, different diagnosis matching strategies, different nurse reviewers, and different physician reviewers were explored. A cost-benefit estimation was also performed. It was concluded that Iliad review was effective and efficient in detecting diagnostic errors causing quality of care problems in Medicare hospitalizations.
This dissertation is dedicated to my parents, Mr. Lau Soon See and Mdm. Tan Soek Kee,
as a gesture of my love and respect.

Thank you for cheerfully supporting my exploration in medical informatics,
in the face of conventional thinking,
after having seen me through medical school.

Especially, thank you for being the best parents I could have,
for teaching me to be happy, to like myself, to enjoy life.

May you have many more years of putting up with me.
# CONTENTS

ABSTRACT ........................................................................................................ iv

LIST OF TABLES ............................................................................................ ix

LIST OF FIGURES .......................................................................................... x

ACKNOWLEDGMENTS ................................................................................ xi

INTRODUCTION ............................................................................................ 1

Scope of Dissertation Project ........................................................................ 1
Current Problems in Health Care .................................................................. 1
Medicare: Government as Direct Payer of Health Care .............................. 2
Medicare Inpatient Quality Review / Intervention ..................................... 5
Quality of Care in Medicare Hospitalizations ............................................. 10
Research hypothesis .................................................................................. 12

LITERATURE REVIEW ................................................................................ 13

Quality of Care Review .............................................................................. 13
Occurrence of Diagnostic Errors ............................................................... 17
Impact of Diagnostic Errors ........................................................................ 18
Using a Computer to Detect Diagnostic Errors .......................................... 19

METHODS .................................................................................................... 26

Experiment Design ..................................................................................... 26
Iliad Review ................................................................................................. 27
Experiment I: Comparing Iliad and UPRO Reviews .................................. 30
Experiment II: Comparing Iliad and UCDS Reviews .................................. 34
Protection of Human Subjects .................................................................... 38

RESULTS ..................................................................................................... 39

Experiment I: Comparing Iliad and UPRO Reviews .................................. 39
Experiment II: Comparing Iliad and UCDS Reviews .................................. 46

DISCUSSION ............................................................................................... 58

Research Hypothesis .................................................................................. 58
Diagnostic Errors Found by Iliad Review .................................................. 58
The Iliad Review Process ............................................................................ 62
Cost-Benefit Estimation of Iliad Review .................................................... 67
LIST OF TABLES

1. The Health Care Financing Administration (HCFA) selection criteria for Medicare inpatient review ............................................................... 7
2. The HCFA generic quality screens for Medicare inpatient review by nonphysicians ............................................................................ 8
3. The HCFA severity level classification and weighted score used by PROs .......................................................................................... 9
4. Experiment I: results of UPRO and Iliad reviews ................................................. 40
5. Experiment I: quality problems found in the URPO review ............................. 41
6. Experiment I: quality problems found in the Iliad review ............................... 41
7. Experiment I: severity of missed diagnoses and unlikely diagnoses ............... 42
8. Experiment I: comparing the results of UPRO and Iliad reviews .................... 42
9. Experiment I: comparing the quality problems found by both Iliad and UPRO reviews in the same cases ................................................. 43
10. Experiment I: combined UPRO and Iliad review and cost estimation .......... 44
11. Experiment II: the ten most common diagnoses in the cases used .................. 46
12. Experiment II: results of UCDS and Iliad reviews ....................................... 47
13. Experiment II: generalization of Iliad review results to test population ........ 48
14. Experiment II: quality problems found in the UCDS review ....................... 49
15. Experiment II: diagnostic errors found in the UCDS review ....................... 50
16. Experiment II: quality problems found in the Iliad review ......................... 50
17. Experiment II: severity of missed diagnoses and unlikely diagnoses .............. 51
18. Experiment II: combined UPRO and Iliad review and cost estimation .......... 52
19. Effect of different threshold criteria on DER and TPR of Iliad Review ........ 54
20. Comparing the Iliad reviews in the two experiments ................................. 60
LIST OF FIGURES

1. U.S. health care spending as a percentage of the GNP, 1983 to 1993 ..............2

2. Comparing health care spending, infant mortality and life expectancy of various developed countries ..........................................................3

3. Assessment Strategy. Diagnostic and therapeutic outcomes routinely assessed; related process variables assessed only if outcomes do not meet accepted standards .................................................................16

4. Sample size calculation for Experiment I: comparing Iliad and UPRO reviews .................................................................32

5. Iliad review DER and TPR for different flagging thresholds .................54

6. Inter-rater reliability between different nurse reviewers in data entry into Iliad .................................................................56

7. Inter-rater reliability between different gold standard physician reviewers .................................................................57
ACKNOWLEDGMENTS

The past five years have been gloriously fun and exciting, and many people helped to make them so.

I would like to thank my family, particularly my parents, Mr. Lau Soon See and Mdm. Tan Soek Kee, my father-in-law, Mr. Lam Wah Kee, and my sister-in-law, Mdm. Lam Siew Kuen, for their limitless love and support. Although I am not able to verbalize my gratitude, it will always be there.

I would like to thank the faculty and staff of the Medical Informatics department for a wonderful learning experience. I shall always remember Dr. Reed Gardner’s advice to think "so what". Special thanks to Dinny Abaunza for her efforts on my behalf, without which my journey in Medical Informatics would not have begun.

I would like to thank my dissertation committee for their help and advice. Dr. Homer Warner, my advisor, allowed me great freedom in exploration while keeping me on the right path. Dr. Michael Lincoln was always willing to take on extra responsibilities by sacrificing his already meagre rest. Dr. Lloyd Poulsen, the medical director of the Utah Peer Review Organization, showed me how to persevere cheerfully with one’s belief and work. Dr. Allan Pryor, ever insightful, helped me see the big picture on many occasions. Dr. John Williamson, with his boundless enthusiasm, never failed to encourage me. Dr. Peter Haug, the nonvoting member of my committee, helped me make many decisions regarding my dissertation research.

I would like to thank the staff of the Utah Peer Review Organization for their patience and help during the entire project. Thanks also go to Dr. David Bjorkman, Dr. Bruce Bray, Dr. Peter Haug, Dr. Michael Lincoln, Dr. Attilio Renzetti and Dr. Frank Tyler (deceased) who performed case review in the first experiment, and the Utah Peer Review
Organization's nurses and physicians who participated in the second experiment. I would also like to thank Dr. Charles Turner for his help on experimental design and statistical analysis. I must thank the Agency for Health Care Policy and Research for granting me a dissertation grant, 1-R03-HS06947-01, which supported part of the research, particularly the second experiment.

I would like to thank Dr. Allan Pryor and Dr. Peter Haug for the additional training I received after the completion of my dissertation project. Thanks also to Dr. Mark Skolnick for his words of encouragement.

Last but certainly not least, I would like to thank my friends, particularly Derick and Lisa Albright, and Roberto and Beatriz Rocha. We have had so much joy together. I am honored.

To all the people who have enriched my life: I've had a wonderful time. Thank you.
INTRODUCTION

Scope of Dissertation Project

This dissertation evaluated a diagnostic expert system in internal medicine as a quality screen for diagnostic errors. It studied the use of a diagnostic expert system to detect potential diagnostic errors that might have given rise to quality problems in patient management. The role of such a quality review tool in the current and future review activities employed by the federal government, for the Medicare population, was explored.

Current Problems in Health Care

Rising Cost of Health Care

The cost of health care in America has long been a topic of national concern. Health care expenditures have been rising (see Figure 1), despite increasing cost-containment efforts on the part of government and industry. $809 billion was spent on health care in 1992, and the 1993 health care costs have been estimated to be over $900 billion. The health care system in has become so expensive that 35 million people could not afford health insurance.\(^1\)-\(^6\)

Quality of Health Care

The large increase in health care spending has naturally led to an increasing attention on health care quality, particularly in light of cost-containment efforts. In 1990, the United States spent more of its Gross National Product (GNP) on health care than any other developed nations, yet lagged in health status indicators such as life expectancy or infant mortality rates (see Figure 2). Although these crude measures were affected by socioeconomic factors and thus not accurate indicators of quality by themselves, they brought attention to the need for good health care in the United States.\(^3\)-\(^4\)
Furthermore, the medical literature showed that there were indeed problems. Chassin, Wennberg and others documented a considerable proportion of inappropriate use of procedures such as coronary angiography.\textsuperscript{7-15} There was also evidence linking inappropriate use to variations in utilization rates.\textsuperscript{7} The Health Care Financing Administration (HCFA) has found large differences in hospital-based mortality rates.\textsuperscript{16} Other studies showed that variations in appropriateness of care resulted in difference in risk-adjusted outcomes.\textsuperscript{17-20}

**Medicare: Government as Direct Payer of Health Care**

In 1965, Congress enacted Title 18 of the Social Security Act - health insurance for the aged, or Medicare. Under Medicare, the federal government provided hospitalization and supplementary medical insurance benefits to senior citizens 65 years or older. At the same time, the less extensive Medicaid program was enacted with Title 19 - grants to states for medical assistance programs, or Medicaid. Under Medicaid, the federal government gave state governments some financial help to provide medical care to poor citizens.\textsuperscript{1,2,16}
Currently, Medicare has expanded its coverage to include citizens, under the age of 65, granted eligibility by the Social Security Administration because of disability, kidney failure or kidney transplant. These made up 4% of the Medicare population.\textsuperscript{6,21} The remaining 30 million Medicare recipients, citizens 65 years or older, constituted 12% of the United States population in 1987. By the year 2030, the proportion is expected to increase to 21% (67 million).\textsuperscript{22}
Reimbursement: PPS and DRG

Since the enactment of Medicare and Medicaid, federal spending on health care has increased dramatically. Medicare reimbursement alone accounted for 40% of the federal health care budget. Medicare started by paying for all reasonable medical services billed, called fee-for-service. Soon, the need for fiscal restraint was apparent. Earlier cost-containment programs included the 1950s' foundations for medical care, the Experimental Medical Care Review Organizations (1970-1975), and the Professional Standards Review Organizations (PSROs), established by Congress in 1972. They were all local, not-for-profit, physician-sponsored organizations which concentrated on reviewing the necessity and quality of services before reimbursement was granted. Despite some success in quality improvement, these programs were believed to have failed in cost-containment.

After PSROs were phased out in 1983, Congress enacted the present Prospective Payment System (PPS) of Medicare reimbursement in the same year. Under PPS, acute-care hospitals are reimbursed a fixed amount according to which of nearly 500 Diagnostically Related Group (DRG) a Medicare patient belongs to. A DRG grouped related disease conditions requiring similar amount of services and care, after adjusting for non-medical factors such as area wage rates. The hospital would thus be paid according to the type and severity of the case, regardless of the actual cost of care. Physician and outpatient services continued to be reimbursed as fee-for-service.

Because 40% of all health care spending went to hospitals, PPS was designed to be an incentive for hospitals to provide only the necessary services consistent with good care. PPS has been successful in slowing the rise in Medicare spending. From 1978 to 1983, Medicare expenditure increased 128%. In the five years after PPS was established (in 1983), Medicare spending only rose 54%. Medicare is currently managed by the Department of Health and Human Services through its Health Care Financing
Administration (HCFA). Reimbursement is through state fiscal intermediaries, such as Blue Cross and Blue Shield in Utah.²,6,²⁹

Quality Assurance: HCFA and PROs

In 1982, the Tax Equity and Fiscal Responsibility Act established the Utilization and Quality Control Peer Review Organization program (commonly known as the PRO program).²,6,³⁰ Because of the prevailing concern that PPS/DRG reimbursement would lead to poorer quality of care, the PRO program was designed "to ensure that services provided to Medicare beneficiaries meet professionally recognized standards of health care, are medically necessary, and are provided in the most appropriate setting."³⁰ HCFA contracted with physician-created or sponsored organizations, called Peer Review Organizations (PROs), in each state. Although each PRO is allowed some autonomy and flexibility in taking into account regional differences, HCFA established guidelines by which PROs carried out their quality assurance functions. The details are specified in a HCFA document called the Scope of Work (SOW).³⁰-³² The third SOW period started in 1989.⁶ In Utah, Medicare quality review is performed by the Utah Peer Review Organization (UPRO), a not-for-profit corporation created and sponsored by Utah physicians.³³

Medicare Inpatient Quality Review / Intervention

Medicare hospitalizations accounted for 34% of all acute care admissions in America. In the first five years since PPS started, there were nearly 45 million hospital discharges involving Medicare beneficiaries.³⁴ PROs have been given the responsibility of monitoring the care received by these Medicare beneficiaries. PROs are required by HCFA to review:

1. quality of care, including appropriateness of discharge,
2. admission necessity,
3. procedure necessity and coverage, and
4. DRG validity.\textsuperscript{30-32}

As of February, 1989, the PRO system reviewed approximately seven million (16\%) of the 45 million Medicare cases.\textsuperscript{34} Because Medicare patients accounted for a large proportion of hospitalizations, and the PRO quality review is typical of the chart review activities widespread in hospitals,\textsuperscript{6} this dissertation focused on the PRO quality of care review for Medicare hospitalizations.

**PRO Quality of Care Review**

The goal of quality review is to improve health care delivery through the monitoring and analysis of patient management strategies. The underlying principle is that detection of a quality of care problem would lead to feedback to the care giver and thus improvement in future patient care.\textsuperscript{35-46}

Based on this concept, PROs performed retrospective chart review for Medicare admissions. Ideally, the quality review would be performed exclusively by expert physicians.\textsuperscript{32} However, the high cost of physician peer review and the great volume of Medicare cases prohibit a physician review of all cases. Therefore, PROs have used HCFA criteria (see Table 1) to sample a subset of Medicare cases for a two-stage method of review known as generic screening. In the first stage, medical charts are reviewed by nonphysician reviewers to identify adverse events that are thought to indicate poor quality of care. These adverse events are defined and identified by generic quality screens (criteria) that are applicable across all diseases.\textsuperscript{6,30,33,47} An example of a generic screen would be "occurrence of a severe adverse drug reaction." Charts with identified adverse events, indicating potential quality problems, would be referred to a physician for the second stage of review, in which the quality problem is confirmed or denied.\textsuperscript{6,47}

Generic screening is the method chosen by HCFA for PROs to review Medicare hospitalizations. The sampled Medicare inpatients charts are subjected to a preliminary screening by nonphysician reviewers. These first-line reviewers, usually nurses or record
Table 1. The HCFA selection criteria for Medicare inpatient review.

- 3% random sample of all hospitalizations.
- 25% random sample of cost and length of stay outliers.
- 25% random sample of readmissions within 31 days of discharge.
- 20% random sample of intervening care.
- 15% random sample of specialty hospital admissions.
- 5% random sample of preauthorized procedures.
- All DRG adjustments that increased payment.
- All cases referred by HCFA or fiscal intermediaries.
- All non-covered admissions later requiring coverage.
- All diagnoses usually not requiring hospitalization.
- Variable sampling of diagnoses or Diagnostically Related Groups (DRG) previously identified as prone to quality of care problem or coding error.
- Variable sampling of transfers, depending on type of facility and diagnosis.
- Intensified review cases.

Professionals trained in the review procedure, applied generic quality screens provided by HCFA (see Table 2).\textsuperscript{30} They can classify a case as being free from quality of care problems. However, they can also flag a chart, that is, identify potential quality problems. These charts are then referred to a PRO consultant physician for peer review. Specific questions, arising from the quality problems, would be asked by the nonphysician reviewer. For example, one HCFA screen required the nonphysician reviewer to check if the patient’s temperature was normal in the 24 hours preceding discharge. If not, the case would be referred to a physician reviewer with the question "is there a quality problem, as the patient still had a fever on the day of discharge?" Although the physician reviewer was not required to use the HCFA guidelines, he was expected to answer the question after he had reviewed the case.\textsuperscript{30-34}
Table 2. The HCFA generic quality screens for Medicare inpatient review by nonphysicians.

1. Adequacy of discharge planning:
   - documentation
   - follow-up

2. Medical stability at discharge:
   - blood pressure
   - temperature
   - pulse
   - other abnormal findings
   - intravenous drugs and fluids
   - wound discharge

3. Death in hospital:
   - during or following surgery
   - following return to intensive care within 24 hours of transferring out
   - when unexpected


5. Unscheduled return to surgery.

6. Trauma suffered in hospital:
   - unplanned surgery
   - fall
   - complication of anesthesia
   - transfusion error
   - decubitus ulcer
   - medication error
   - inappropriate care or lack of care

Quality Problem Severity Level and Score

Each quality problem confirmed by the physician review is rated for severity by the national PRO severity level classification established by HCFA. Each confirmed quality problem would also be assigned a weighted severity score (see Table 3). If a case had more than one quality problem, the highest score would be assigned to the case.30,32,34

PRO Intervention Program

When a quality problem is found by the physician review, the PRO notifies the hospital or attending physician responsible for the care of the patient. Further information is requested from the caregiver. If additional information is provided, the case is reviewed
Table 3. The HCFA severity level classification and weighted score used by PROs.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Definition</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care appropriate:</td>
<td>No quality problem.</td>
<td>None</td>
</tr>
<tr>
<td>Level I quality problem:</td>
<td>No potential for significant adverse effects on the patient.</td>
<td>1</td>
</tr>
<tr>
<td>Level II quality problem:</td>
<td>With potential for significant adverse effects on the patient.</td>
<td>5</td>
</tr>
<tr>
<td>Level III quality problem:</td>
<td>With significant adverse effects on the patient.</td>
<td>25</td>
</tr>
</tbody>
</table>

Significant adverse effect was defined as:
1. Unnecessarily prolonged treatment, complications, or readmissions.
2. Patient management which resulted in anatomical or physiological impairment, disability or death.

by a second physician reviewer, before the quality problem is finally confirmed and a severity level and weighted score assigned.30,33,34

The weighted scoring system is used by PROs to generate a quality of care profile. Based on the number of points accumulated in a quarter, one of the following corrective interventions may be implemented:

- Notification - if 3 points are reached. The aim is to promote awareness of problem and to encourage self-assessment and improvement.

- Education - if 10 points are reached. Depending on the quality problem, literature or reference material may be provided, or completion of a written paper or a Continuing Medical Education (CME) course may be required.

- Intensification - if 15 points are reached. An intensified case selection and review is carried out on the hospital or physician's patients.

- Other interventions - if 20 points are reached. A specialty-oriented quality review committee is formed to consider:
  1. referral to a hospital quality assurance committee,
  2. referral to the State Licensing Board,
3. sanction recommendation to the Department of Health and Human Services.\textsuperscript{30,33,34}

In the review period of April, 1989 to March, 1990, the Utah Peer Review Organization (UPRO) increased the review selection of four Utah hospitals because of questioned admitting patterns. The problems were resolved. Also, two hospitals and 18 physicians were given intensified reviews because of quality of care problems.\textsuperscript{33}

**Quality of Care in Medicare Hospitalizations**

**PRO Review Underdetected Quality Problems**

From April, 1989 to March, 1990, UPRO reviewed 8,648 Medicare cases, or about 22\% of the total Medicare claims (39,850 cases) received during that period in Utah. The 8,648 Medicare cases first underwent the nonphysician review, which at UPRO is carried out by trained nurse reviewers. The nurses referred 18\% of the reviewed cases to physician reviewers, who ultimately confirmed only 96 cases to have quality problems. These 96 cases represented 1.11\% of the initial 8,648 cases reviewed by UPRO.\textsuperscript{33} This was in line with the national PRO average quality problem rate, which was 1.57\%.\textsuperscript{48}

Studies have indicated that the current PRO quality review activities are at best only moderately effective.\textsuperscript{49-52} For instance, Rubin et al. compared PRO review results with their own independent quality assessments. They concluded that PRO review detected only 11\% of the cases containing quality problems. Based on their research, the authors estimated that the true quality problem rate in Medicare records was in the region of 18\%,\textsuperscript{49} rather than the 1.57\% reported as the national average.\textsuperscript{48} In another study, the Prospective Payment Assessment Commission (ProPAC, a Medicare commission established by Congress to recommend changes in DRG classification and reimbursement)\textsuperscript{2,6} reviewed more than 3,000 records using the PRO generic quality screens. The study found quality problems in 5\% of the cases. In addition, ProPAC estimated that an equivalent number of problem cases were not detected by the screens.\textsuperscript{50} Studies using other evaluation methods had also shown much higher rates of quality problems,\textsuperscript{6,9,53} Therefore, the current PRO
review procedure might have seriously failed to detect some important quality problems in Medicare cases.

**PRO Review Failed to Detect Diagnostic Errors**

Based on a scrutiny of the current PRO review procedure, this dissertation hypothesized that some of the undetected quality problems might have arisen from diagnostic errors. Because physicians keyed their management strategies to diagnosis, diagnostic errors could result in a seriously flawed patient management strategy. Diagnostic errors could lead to delayed or inappropriate investigations or treatment,\(^8,9\) in turn resulting in adverse outcomes.\(^{54-57}\) Because the present PRO review process began with the identification of probable quality problems by nonphysician reviewers, mostly nurses, diagnostic quality problems might not be apparent, unless the reviewer realized that the diagnosis is incorrect. Only then would these quality problems be referred to physician reviewers. Nurses were unlikely to be specifically trained in diagnosis, and the HCFA guidelines used in the nurse review focused almost entirely on therapeutic and management errors (see Table 2). Thus, a proportion of the quality problems not detected by nurse review were likely due to diagnostic errors. Rubin et al., in the study mentioned earlier, found that the PRO nonphysician screening failed to flag and refer two out of three cases judged to have quality problems by an independent review.\(^{49}\) As another example, UPRO nurses questioned the diagnosis on only about a dozen charts out of the 8,648 charts reviewed in the period of April, 1989 to March, 1990 (verbal communication with Dr. A. Lloyd Poulsen, medical director of UPRO). Medical research indicates that diagnostic errors were much more widespread than this low UPRO rate suggested, and that diagnostic errors, through their influence on subsequent patient management, could adversely affect patient outcome.\(^8,9,54-60\)

It could be argued that since diagnostic errors result in a flawed patient management strategy, such cases were likely to contain quality problems, some of which should be detectable by the nurse review, even without recognition of the diagnostic error and the
other quality problems consequent to it. Thus, when the problem case was referred to and reviewed by the physician reviewer, the diagnostic error and the additional quality problems not detected by the nurse should then be found by the physician. However, a physician referral always came with questions raised by the nurse, which the physician was required to answer. Therefore, the attention of the physician reviewer was likely focused on specific quality concerns, rather than on performing a broad quality review, including the diagnosis, of the case. An examination of UPRO review results lent support to this view, as among the 86 problem cases found in a 12-month period, none had the diagnosis questioned by either the nurse or the physician reviewer.

Research Hypothesis

This dissertation hypothesized that a diagnostic expert system - a computer program designed to suggest medical diagnosis - could be used to help the nurse reviewer detect potential diagnostic errors that would not have been found by the current review method. These potential problem cases could then be referred for physician review with the appropriate question on diagnosis.
LITERATURE REVIEW

Quality of Care Review

Definition of Quality of Care

There have been numerous definitions of quality of care in the medical literature. In 1933, Lee and Jones defined "good medical care" as "the kind of medicine practiced and taught by the recognized leaders of the medical profession at a given time or period of social, cultural, and professional development in a community or population group." This definition is intuitive to most medical practitioners and is the principle underlying physician peer review activities, such as those adopted by the PROs.

Esselstyn, in 1958, offered a broader definition of quality care. He stated that "standards of quality of care should be based on the degree to which this care is available, acceptable, comprehensive, continuous, and documented, as well as on the extent to which adequate therapy is based on an accurate diagnosis and not on symptomatology." This definition added the structural consideration of availability of, or access to, health care services to the process considerations of care comprehensiveness and acceptability, or standard. The terms structure and process were proposed by Donabedian in 1966, referring to the settings in which care took place and the medical care process, respectively.

Esselstyn's definition also highlighted the importance of care continuity (follow-up) and documentation. Most importantly, Esselstyn raised the idea that medical management should be effective and efficient - basing adequate treatment on accurate diagnosis. This has evolved into today's focus on health outcomes, and the recent Institute of Medicine (IOM) definition of "quality of care" as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with
current professional knowledge.6 The IOM definition of quality of care focused attention on the relationship between the health care process and the resultant patient outcome. It also reflected the growing concern that much of what is commonly practiced in medicine had not been proven to be efficacious or even effective.

From these examples of quality of care definition, it can be seen that judgment on quality of care is dependent on the individual's perception of what constitutes good care, and that in turn is a product of the health care system. The concept of quality care will continue to be defined and shaped by changing socioeconomic forces as well as scientific and technological advances.

**Assessment of Quality of Care**

Quality assurance has been defined as "a formal and systematic exercise in identifying problems in medical care delivery, designing activities to overcome the problems, and carrying out follow-up monitoring to ensure that no new problems have been introduced and that corrective steps have been effective."35 Classically, quality assurance is viewed as a cyclical set of activities designed to improve the quality of care. Identification of a quality problem is followed by corrective activities, then monitoring to ensure that the problem has been solved and for new problems, and so on in an ever-continuing cycle. The term quality assessment referred to the measurement of the quality of care, the starting point in quality assurance. Quality assessment through reviewing the medical care for quality problems has been called quality review.6 The medical literature has documented many different approaches, methods and programs for quality assurance and assessment.23,35-46

In 1966, Donabedian proposed three approaches for assessing quality of care: structure, process and outcome.36 Structure referred to the capacity to provide care. Examples of structural variables that could be assessed include physical (size, for example) and organizational characteristics (such as staff-to-patient ratio) of the settings in which care is given. Structural attributes, when applied to health care professionals, include factors
such as age and certification. Process referred to the medical care process of diagnosis and treatment. Assessing process, for instance, by retrospectively reviewing the care documented in the patient chart, would produce a direct judgement of care quality. This is the most frequently chosen approach to quality assessment.\(^6\) The PRO quality review focused on the process of care by reviewing the medical record, using the generic screening method proposed by Rutstein in 1976.\(^47\) Outcome referred to the state of health or well-being, of the patient or population, as an end result of medical care.

Donabedian's classic paradigm has been widely adopted as the "unifying conceptual framework for quality measurement and assurance."\(^6\) Other researchers have proposed different approaches or added to Donabedian's tenets. De Geyndt suggested five approaches for assessing quality of care: content, process, structure, outcome, and impact. He proposed splitting off content from process of care (considered together under process by Donabedian) for conceptual clarity, and added the assessment of impact of care on the whole population.\(^37\) Densen divided quality assessment methods into two groups. The medical audit method evaluated the process of care against established standards. The end result method judged quality as the ability to increase physical and social well-being, analogous to outcome assessment.\(^38\) Approaching from the standpoint of how rather than what to evaluate, Peterson proposed three types of quality assessments: end result studies, direct observations and chart reviews.\(^39\) The last, retrospective medical record review, is the method used by PROs in reviewing Medicare hospitalizations.

The 1980s has seen a dramatic increase of interest in outcomes research.\(^64-69\) John Williamson proposed the Health Accounting approach to quality assurance, a strategy that relates outcome and process assessment.\(^68,69\) In his approach, diagnostic and therapeutic outcomes are routinely assessed, and when "achievable benefits" are not achieved, the process of care (diagnostic or therapeutic) is examined for the cause of poor outcome (see Figure 3). In order of priority for evaluation, the four elements of this strategy are
diagnostic outcomes, diagnostic process, therapeutic outcomes, and therapeutic process. In the PRO review, some attempt to look at therapeutic outcome is seen in the severity level classification for the quality problems found. Diagnostic outcome is virtually ignored, and the diagnostic process received very little attention. The main focus of the PRO review is on therapeutic process, the least important on Dr. Williamson’s list.

In general, quality assessment has been the most emphasized component of traditional quality assurance activities, which functioned by weeding out bad apples. Recent years have seen the emergence of a contrasting quality assurance approach, called Continuous Quality Improvement (CQI) or Total Quality Management (TQM). Adopted from industry, the CQI/TQM approach also involved a cyclical series of steps. The desirable end-product would first be identified. The process of care would then be redesigned to meet the planned objectives. The new process would be monitored and assessed, and areas of possible improvement (desirable end-product) again identified. The CQI/TQM approach aimed to improve the performance of every person, process or system involved in patient care, in contrast to traditional quality assurance approaches which improved the overall standard of care by removing outliers.
There has been a long tradition of quality assurance in the medical community. Different approaches and numerous methods have been developed. It seemed clear that, whichever the approach, quality assessment is an integral part of quality assurance. By measuring the quality of care rendered, quality assessment provided a platform for improvement efforts. This dissertation focused on the PRO quality assessment activities - the retrospective chart review to detect quality problems.

**Occurrence of Diagnostic Errors**

Although the PRO review process might have been unable to detect diagnostic errors, other researchers have quantified their occurrence. A few recent examples are reported here. Mercer and Talbot compared the clinical diagnoses of 400 patients who died in hospital to the pathological findings discovered at autopsy. The authors found confirmation of the clinical diagnoses in only 47% of cases. Both underdiagnosis and overdiagnosis were present. Potentially treatable disease was missed in 13% of patients.\(^\text{58}\) Anderson et al. analyzed over five decades of studies of more than 50,000 autopsies. They found that the clinical accuracy of a positive diagnosis ranged from 60% to 96%, depending on the diagnosis. For carcinoma of the lung, the accuracy of a positive diagnosis was only 86%. It was possible that patients who died and received autopsies suffered from a greater proportion of diagnostic errors. Nevertheless, the presence of diagnostic errors were indisputable. In the case of lung cancer, the authors pointed out that detection rates in autopsied populations were similar to overall population detection rates reported elsewhere.\(^\text{59}\)

Weinberg reviewed the charts of 146 surviving patients with diagnosis in one of five DRGs in the area of internal medicine. He found that the proper diagnosis was either delayed or missed in 10% of the cases. Additionally, 38% of the cases were identified as showing inadequate generation of initial diagnostic hypotheses (following the history and physical examination). Furthermore, 19% of cases showed insufficient diagnostic data acquisition, and 32% showed inattention to or misinterpretation of diagnostic cues.\(^\text{60}\)
Impact of Diagnostic Errors

Studies have shown that diagnostic errors, through their influence on subsequent patient management, directly and adversely affected health outcome. Williamson et al. assessed the quality of ambulatory care provided in two rural and one tertiary medical clinics. The authors suggested that avoidable disability was related to significant clinical errors. Among 71 patients, 47 were judged to have received adequate care, 10 questionable care, 9 inadequate care and 5 unnecessary care. Needed diagnostic investigations were not ordered in 58% of patients in the rural clinics and 50% of those in the tertiary clinics. A notable medication error was found in 53% of rural clinic patients and 35% of tertiary clinic patients. Many problems were directly keyed to erroneous diagnoses. Overall, Williamson and his co-workers judged that only 58% of rural patients and 84% of tertiary clinic patients received good care.54

Other researchers have corroborated Williamson's observations. Kassirer and Kopelman analyzed 40 transcripts of cases with known diagnostic errors. They demonstrated that these diagnostic errors produced serious morbidity, such as unnecessary loss of a kidney and delayed chemotherapy for cancer.55 Pelberg reviewed 64 cases of acute myocardial infarction in which the diagnosis was initially missed (and another diagnosis wrongly assigned). These misdiagnosed patients experienced an 83% mortality rate, compared to an expected mortality rate of 25%.56,57

When physicians committed diagnostic errors, they often inappropriately ordered expensive and potentially dangerous procedures or treatment. This inappropriate management behavior was demonstrated in a study by Chassin et al.. The authors performed a retrospective chart review of 1,677 coronary angiography and 1,585 upper gastrointestinal endoscopy. They found inadequate diagnostic evidence to support the use of coronary angiography in 17% of cases. A similar rate was found for upper gastrointestinal endoscopy.8 Brook et al. analyzed medical literature to assess the appropriateness of acute medical care for the elderly. The authors found 10 studies
reporting inappropriate diagnostic and treatment procedures in 3% to as many as 75% of cases. For instance, one study reported that 14% of coronary bypass surgeries in Medicare patients were not indicated. Another study reported that cardiac pacing was inappropriately ordered in 20% of Medicare cases.9

A survey of the medical literature thus led to the conclusion that diagnostic errors were more common than the PRO review results suggested. The General Accounting Office has found that allegations of diagnostic errors constituted a quarter of all malpractice claims.71 Diagnostic errors could give rise to important quality problems that resulted in adverse outcomes. Unfortunately, these quality problems might not be detectable by current PRO review procedures.

Using a Computer to Detect Diagnostic Errors

The detection of quality problems is the prerequisite to feedback and improvement in health care delivery.23,35–46 Therefore, the failure of PRO review to detect diagnostic errors is an important flaw. A possible solution would be to have physicians perform a general review of a case's diagnosis and management. However, physician review is expensive, and the time-consuming general review would be even more so, especially with the large number of cases to be reviewed. Increasing the problem detection rate by solely using general physician peer review, without the focusing influence of the preliminary nurse review, is thus unlikely to be practical. A more practical approach would be to find a tool to help the nurse reviewer detect potential diagnostic errors, much like the HCFA guidelines helped in the detection of management problems. The cases flagged for diagnostic errors could then be referred to physician reviewers with the appropriate question on diagnosis.

To help the nurse detect diagnostic errors, the quality review tool should be capable of making accurate diagnosis over a large number of disease conditions. In addition, it must be able to provide patient-specific diagnosis, or it would be only as useful as a
textbook. A diagnostic expert system in medicine appeared to be promising as such a quality review tool.

An expert system is a computer program designed to solve particular problems in a particular domain. It has two essential components, the knowledge base and the inference engine. The knowledge base contains the knowledge of a human expert in a database. The process of obtaining the knowledge from the human expert and putting it in a form usable by the computer is called knowledge engineering. The knowledge contained in the knowledge base drives the logic embedded in the inference engine to make such suggestions as the expert would make. A diagnostic expert system thus would suggest diagnosis in a medical area. Because a diagnostic expert system is designed to base its diagnostic suggestions on patient findings, it can be used to provide patient-specific diagnosis. A diagnostic expert system thus might serve the function of alerting the nurse reviewer to the presence of a diagnostic error.

Computers in Quality Assessment

Computers have been used for medical quality assessment in several ways. One application was the maintenance and analysis of large databases containing administrative or claims data. Roos et al. have provided a theoretical discussion on the quality assessment function of such databases. Data could be analyzed for mortality and morbidity, and other outcome studies. Recently, Iezzoni et al. developed a computerized algorithm to screen hospital discharge abstract data, collected in state databases, for indicators of potential quality problems. The computerized screens were designed to suggest hospitals that required a more intensified review of their care. The screens were evaluated by comparing their error flags with recommendations made by physicians. The computerized screens found 93% of the cases picked by physicians as meriting quality review. However, the screens classified as problem-free only 64% of the cases believed by the physicians not to warrant a review. Because of the nature of the data used in the study, no quality review
was performed. Thus, whether the screens would be efficient in targeting hospitals with more quality problems remained unknown.87

Databases have also been analyzed for resource utilization and other quality concerns relating to use of services.88 These applications have been evaluated in various specialty areas such as hemodialysis,89,90 emergency room chart audits,91 blood transfusion audits,92,93 and drug-prescribing behavior modification.94-96 Chae described a computer-based quality assurance system at a large teaching hospital.97 This system provided a more comprehensive information base to establish effectiveness, efficiency and appropriateness standards. However, the system was confined to a single hospital and required a unique mainframe computer and expensive customized software.

The above examples used computers in a retrospective evaluation of care. Meanwhile, computers have increasingly been used, in a prospective manner, to help clinicians deliver better care. Such systems are also called decision-support systems,74 as they are designed to help in the decision-making process. Although these systems are not specifically used in retrospective quality assessment or review, they served quality assurance functions. There have been many examples of computer systems that provided physician suggestions, reminder or alerts.98-108,112-136 For illustration, some of the quality assurance functions of the HELP hospital information system, at the LDS Hospital in Salt Lake City, is mentioned here.

The HELP system is a hospital information system developed by the University of Utah.109-112 A hospital information system is designed to help in the management of information needed by health care professionals. Hospital information systems are in widespread use, but the majority are administrative in function, with limited clinical applications.74 The HELP system is unusual in that it provided clinical decision-support. An example would be a variety of alerts that could be communicated in real-time to physicians, thus directly affecting patient care.112-136 HELP monitored prescriptions for adverse drug-drug, drug-laboratory, drug-diet and drug-allergy interactions, to be reported
to the prescribing physician.\textsuperscript{112-115} Abnormal laboratory findings would trigger the generation of laboratory alerts.\textsuperscript{112,113,116,117} The HELP system also surveyed the use of perioperative antibiotics and the occurrence of hospital-acquired infections.\textsuperscript{112,113,118-128} These alerts have been found to result in improvement such as increased appropriateness of antibiotic usage. Other quality assurance functions provided by the HELP system include respiratory therapy monitoring, blood ordering audit and adverse drug event monitoring.\textsuperscript{112,113,129-136}

There have been some studies that used computers specifically for quality review — a retrospective assessment of quality of care. Stewart et al. described a computerized quality assurance system to supplement a full-time quality assurance officer. The system contained quality screening criteria for emergency room case review. The authors found that the number of patient cases referred for investigation of questionable care rose from a preimplementation rate of five patient care errors per month to 35 per month.\textsuperscript{137} In the Netherlands, van der Lei et al. developed a computer system to critique general practitioners' management of hypertensive patients, by obtaining data from a computerized information system. The investigators compared the automated review to physician reviewer, using the consensus judgement of eight review physicians as the gold standard. The computerized review results were closer to the gold standard than those of individual physician reviewers.\textsuperscript{138} A follow-up study found that general practitioners were equally accepting of the critiques generated by the computerized review and those from physician reviewers.\textsuperscript{139}

Responding to increasing pressure for standardization of the PRO nonphysician review, HCFA started, in 1987, to develop a Uniform Clinical Data Set (UCDS). The UCDS is a rule-based expert system which contained, as review criteria in the form of computer algorithms, the present HCFA generic quality screens used in the PRO nonphysician review. It is planned for use by nonphysician reviewers to perform the preliminary screening portion of the current PRO review process. The program would
request for patient data according to the contained review criteria, and the nonphysician
reviewers would enter the data from the patient chart into the program. UCDS would flag a
case for physician review if one of the review criteria is triggered by the data entered.\textsuperscript{140-142}
Because the review criteria contained in UCDS are the same HCFA guidelines in terms of
content, the only advantages expected to be gained by the UCDS review over the current
manual review are that of objectivity and completeness. It is thus reasonable to expect the
UCDS review process to have a similar deficiency as the current PRO review in detecting
diagnostic errors. UPRO is one of seven PROs currently evaluating the UCDS review
process.\textsuperscript{139} HCFA has targeted UCDS to be used in all PROs by 1994.\textsuperscript{141}

With the increasing presence of computers in clinical, as well as administrative
aspects of medical care, more of such quality review applications could be expected in the
future. This dissertation considered the use of a diagnostic expert system, Iliad, in quality
review.

\textbf{A Diagnostic Expert System: Iliad}

Diagnostic expert systems have been a prominent area of research. There were
numerous systems developed for different domains of medicine.\textsuperscript{75-82} Most of these
systems were narrow in scope. For example, Casnet was an expert system designed to
diagnose glaucoma.\textsuperscript{75-82,143,144} Many of these limited domain expert systems had been
shown to be accurate and useful. For instance, deDombal and his co-workers found that a
diagnostic expert system helped junior emergency room staff improve their recognition of
acute abdominal conditions requiring urgent surgery.\textsuperscript{75-82,145-147} However, these restricted
domain expert systems would be of limited use as a general screening tool in quality
review.

There were some diagnostic expert systems capable of recognizing large number of
diseases. The notable ones included QMR, Dxplain and Iliad.\textsuperscript{148-162} All three were
designed for the domain of internal medicine. QMR was developed by Miller and his co-
workers at the University of Pittsburg.\textsuperscript{149-151} It evolved from its predecessor, the Internist
program, and could diagnose one of nearly 600 diseases. The diagnostic accuracy of QM had been evaluated to be in the region of 85%. Dxplain was developed by Barnett et al. at the Massachusetts Institute of Technology. It recognized 2,100 diagnoses. Dxplain had been evaluated at 73% accuracy, and a later study found that its diagnostic performance compared favorably to that of expert physicians. Iliad was developed by Dr. Homer Warner at the University of Utah. It was the diagnostic expert system chosen for evaluation in this dissertation.

Iliad represents the culmination of over 20 years of research into expert systems at the University of Utah Department of Medical Informatics. Iliad is an expert system designed for diagnosis in the field of internal medicine, but the system also includes diagnostic knowledge in certain related fields with which a primary care physician must be familiar. Examples are common conditions in surgery, dermatology, psychiatry, obstetrics, gynecology and orthopedics. Currently, Iliad (version 4.0) recognizes over 6,300 medical findings and 1,350 diagnostic conditions. Iliad consists of a knowledge base and an inference engine. The knowledge base is continually expanded, evaluated, and maintained as part of an ongoing knowledge engineering effort. A recent assessment showed Iliad to have correctly identified the differential diagnosis in over 95% of patient cases (unpublished data). The system is microcomputer-based, running either on an Apple Macintosh (Apple Computer, Inc., Cupertino, CA) computer, or an IBM (IBM Personal Computer, Co., Somers, NY) or compatible machine. User interaction with the Iliad system is via a mouse (point-and-click device) and keyboard.

*Iliad is also a trademark of Applied Medical Informatics, Inc. (Salt Lake City, Utah) for the commercial version. Only those features of Iliad used in this dissertation research has been described. For information on Iliad™, the reader is referred to the Iliad user manual published by Applied Medical Informatics, Inc.*
Iliad as a Quality Review Tool to Detect Diagnostic Errors

In Iliad's consultation mode, medical findings (history, physician examination, laboratory data and procedure findings) were entered sequentially into Iliad. At any stage in this data entry, after each new finding is entered, Iliad could provide a list of differential diagnoses which explained the findings entered thus far. These diagnoses were ranked on the differential list according to the diagnostic certainty (current probability) Iliad assigned to each, given the medical findings that supported each diagnosis.\(^\text{163}\)

This dissertation proposed to use Iliad to alert the nonphysician reviewer to the presence of a potential diagnostic error in the following manner: because Iliad based its diagnostic suggestions on a given set of patient findings, if the nonphysician reviewer entered the clinical findings of a patient case into Iliad, and it generated the same diagnosis as that made by the patient's attending physician, the case was likely to be free from diagnostic errors. On the other hand, if Iliad suggested a diagnosis different from that made by the attending physician, the case could then be referred to the PRO physician reviewer, who would judged whether the attending physician committed a diagnostic error or Iliad flagged a false positive.
METHODS

Experiment Design

Two experiments were planned for this dissertation. The first experiment was designed to determine how Iliad might be used to improve the PRO quality review, typified by that performed in UPRO, by screening for diagnostic errors in the nonphysician review. This first experiment thus evaluated Iliad's performance using data obtained from UPRO's routine Medicare review. A second experiment was planned because, as mentioned earlier, UPRO was evaluating the Uniform Clinical Data Set (UCDS). HCFA had targeted nationwide implementation of UCDS, at least for specific projects in PROs, by 1994. Because UCDS contained the same HCFA nonphysician review guidelines used in the current manual PRO review, it likely ignored diagnostic errors as well. Therefore, the second experiment was designed to evaluate Iliad review against the UCDS review performed by UPRO.

The term Iliad review was used to mean the entire review process of using Iliad to flag potential problem cases containing diagnostic errors, and the subsequent physician review to confirm or deny the quality problems. To maintain descriptive uniformity, the data entry and diagnostic error flagging portion was referred to as Iliad nurse review, analogous to the UPRO nurse review. Iliad's physician review was called gold standard physician review, to differentiate it from the UPRO physician review. The term UPRO review thus included the UPRO nurse review and the UPRO physician review. Similarly, UCDS review included the UCDS nurse review, in which the nurse reviewer used the UCDS program to perform the nonphysician review currently done manually in the UPRO nurse review, and the UCDS physician review.
Because nonphysician reviewers in UPRO were all nurses, the shorter term nurse review was used for the nonphysician review portion of all three review processes.

Apart from the source of data, the two experiments had one major difference: the gold standard physician reviewers (for the Iliad review). The first experiment involved volunteer physicians from the University of Utah School of Medicine, none of whom were on UPRO's panel of physician reviewers. During the second experiment, additional funding became available (AHCPR dissertation grant no. 1-R03-HS06947-01). UPRO review physicians were then invited to participate as gold standard reviewers for the second experiment.

**Iliad Review**

**Iliad Nurse Review: Data Entry**

Iliad's consultation mode was used in the Iliad-aided quality review. Patient findings from history, physical examination and investigative laboratory tests were obtained from the case record. Each case review was started by entering the patient's age and sex, followed by the chief complaint. The chief complaint and other case findings were entered into Iliad by typing a keyword, for instance "abdominal pain." The program used pattern-matching to bring up one or more corresponding data items contained in the knowledge base. For example, typing "abdominal pain" brought up "present history: abdominal pain," "past history: abdominal pain," and "abdominal palpation: tenderness," among others. The chosen data items were confirmed by pointing a screen arrow (via moving the mouse or pressing the arrow keys on the keyboard) and typing or clicking with the mouse on "Y" (for yes), "N" (for no) or "?" (for unknown) for each item. Alternatively, Iliad were asked to present an alphabetically arranged list of all data items contained in its knowledge base, for selection and confirmation as described above. After the chief complaint was entered, Iliad presented a list of differential diagnoses, ranked according to the diagnostic probability Iliad assigned to each diagnosis. Iliad updated this differential list after each new finding was entered. The final differential list, after all
available data were entered for the case, was compared to the attending physician's diagnosis for the case.

The Iliad nurse review data entry in this dissertation was performed by this author, a general practitioner. Because this dissertation planned to evaluate Iliad's use in nonphysician review, care was taken by this author not to impose her medical judgement when obtaining patient data from the charts. In other words, all available data from the charts were used as is without additional interpretation. The data entry for all cases were completed before the diagnostic lists (Iliad's and attending physician's) were tabulated and compared.

**Iliad Nurse Review: Flagging a Diagnostic Error**

A potential diagnostic error was identified when there was a discrepancy between the attending physician's list of diagnoses and the corresponding list provided by Iliad for the case. A discrepancy was defined as one of the following conditions:

- A diagnosis on the attending physician's list did not reach above 20% probability on Iliad's list (considered a potential unlikely diagnosis made by the attending physician).
- A diagnosis exceeding 80% probability on Iliad's list did not appear on the attending physician's list (considered a potential missed diagnosis by the attending physician).

The threshold values of 80% and 20% were chosen based on past experience with Iliad. It was thought that when a diagnosis in Iliad reached a probability of more than 80%, the diagnosis was likely to be truly present; conversely, when a diagnosis in Iliad failed to reach a probability of 20%, the diagnosis was likely to be truly absent. An identified discrepancy between the attending physician's diagnostic list and Iliad's would flag the case as requiring a gold standard physician review to confirm or deny the diagnostic error.
In comparing the attending physician’s diagnosis to Iliad’s, postoperative complications and chronic conditions with no acute manifestation were not included. Because Iliad only used the data obtained during the admission diagnostic workup, it would not diagnose postoperative complications, as well as chronic conditions which were not the cause for admission and did not manifest during the hospitalization. An example of a chronic condition would be hypertension. Iliad would not diagnose hypertension if the condition was known and treated before admission, and the patient remained normotensive during the course of the present illness.

This step of the experiment was also performed by this author, after all data entry was completed.

**Gold Standard Physician Review**

Each of the charts flagged by a discrepancy between the physician’s and Iliad’s diagnostic lists was submitted to a gold standard physician review. Iliad’s diagnoses were not revealed to the gold standard review physician. Instead, specific questions arising from the discrepancy were asked. For instance, if the attending physician diagnosed unstable angina but not acute myocardial infarction, and Iliad diagnosed acute myocardial infarction with a probability of 87%, the gold standard physician reviewer would be asked if the patient could have suffered from an acute myocardial infarction. After the gold standard physician reviewed the case, if he or she replied in the affirmative, he or she would be asked if there was then a quality problem in the management of the case, and if so, to rate the problem severity using the PRO quality classification described earlier (in Table 3). A severity score was assigned to each case according to the severity level of the quality problems, using the PRO weighted scoring system, also described earlier (in Table 3). If a case had more than one quality problem, the highest severity score would be assigned to the case.
Experiment I: Comparing Iliad and UPRO Reviews

Case Selection

UPRO reviewed 8,648 Medicare inpatient cases from April, 1989 to March, 1990. A paper copy of the 1,557 charts flagged by the nurses as containing a potential quality problem was kept in the UPRO. That is, each of these 1,557 charts had been reviewed by a UPRO nurse, found to have contained at least one potential quality problem, and referred to a UPRO review physician. The physician review may or may not have confirmed the flagged problems.

Cases were selected from these 1,557 charts for the first experiment, because it was hypothesized that Iliad would detect a different type of error from that focused on by the UPRO review, and also because charts classified as problem-free by the UPRO nurse review were not kept. It was decided not to choose only the problem cases (confirmed by the UPRO physician review) for the experiment, because the True Positive Rate (TPR, also known as sensitivity) of the Iliad nurse review could then be compared to that of the UPRO nurse review:

\[
\text{TPR of nurse review (Iliad or UPRO)}
= \frac{\text{True positive flags}}{\text{True positive flags} + \text{false positive flags}}
= \frac{\text{Number of cases correctly flagged as having problems}}{\text{Total number of cases flagged as having problems}}
= \frac{\text{Number of problem cases confirmed by physician review}}{\text{Total number of cases referred for physician review}}
\]

The nurse review TPR was an indication of cost-effectiveness and efficiency, since false positive flags resulted in unnecessary physician review costs. UPRO reported the TPR of its nurse review to be 5.5% for the 1,557 charts, resulting in only 86 problem cases confirmed by the UPRO physician review. It was hoped that the Iliad nurse review would produce a higher TPR.

A list of the approximately 1,000 diagnoses contained in Iliad (version 2.03, August 1990) was given to UPRO. From among the 1,557 cases, those containing at least
one of the Iliad diagnoses were selected. Next, the cases with absolutely no diagnostic
work-up information (history, physical examination findings, and laboratory test results) in
the charts were eliminated. These cases usually have had their diagnostic work-up
performed previously, and were admitted for treatment (such as surgery) only. UPRO
provided 242 charts which met the case selection criteria. From among them, 100 charts
were randomly selected for use in the experiment.

Sample Size

The 86 quality problems found by the UPRO review represented 1% of the
Medicare cases reviewed in the 12-month period. However, none of the problems were
diagnostic in nature. Therefore, the diagnostic error rate is 0% in the null hypothesis
population. It was decided to calculate a sample size sufficient to detect a 1% diagnostic
error rate, based on the assumption that the number of diagnostic errors would be on a par
with that of management problems. For the experiment, $\alpha$ was set at 0.05 (for a one-tailed
test, because the diagnostic error rate found by Iliad review could only be 0% or larger),
and power was set at 0.80 (thus $\beta$ set at 0.20). Sample size ($n$) was calculated with the
formula for comparing proportions, given in Figure 4. For the experiment, it was
decided to increase the number slightly, to 100 cases.

Research Hypotheses

Experiment I was designed to evaluate the usefulness of Iliad review by measuring
the number of diagnostic errors found. Cost-effectiveness would be reflected by the TPR
of the Iliad nurse review. The following hypotheses were tested:

IA: Iliad review detected a higher Diagnostic Error Rate ($DER_1$) for the test
population, compared to the Diagnostic Error Rate found by the UPRO review
($DER_0$).

Null hypothesis ($H_0$): $DER_1 \leq DER_0$

Alternative (experiment) hypothesis ($H_1$): $DER_1 > DER_0$
\[ \sqrt{n} = \frac{z_a \sqrt{2R_c(1 - R_c)} - z_B \sqrt{R_t(1 - R_t) + R_c(1 - R_c)}}{R_t - R_c} \]

\[ = \frac{1.65 \sqrt{2 \times 0(1 - 0)} - (-0.84) \sqrt{0.01(1 - 0.01) + 0(1 - 0)}}{0.01 - 0} \]

\[ n = 70 \]

where:  
\( n \) = sample size,  
\( za \) = the one-tailed z value related to the null hypothesis,  
\( zB \) = the one-tailed z value related to the alternative hypothesis,  
\( Rc \) = the error rate in the control population, and  
\( Rt \) = the error rate in the test population.

Figure 4. Sample size calculation for Experiment I: comparing Iliad and UPRO reviews.

IB: Iliad nurse review had a higher TPR (TPR_1), compared to the TPR produced by the UPRO nurse review (TPR_0).

Null hypothesis (H_0): TPR_1 \leq TPR_0

Alternative (experiment) hypothesis (H_1): TPR_1 > TPR_0

Statistical Analysis

The outcomes of the experiment were the Diagnostic Error Rate (DER) found by the Iliad review, compared to that found by the UPRO review, and the TPR of Iliad nurse review, compared to that produced by the UPRO nurse review. The test population was 100 Medicare inpatient cases randomly selected from those flagged by UPRO nurse review in a 12-month period, and containing diagnoses recognized by Iliad. The DERs and TPRs calculated in this experiment applied only to this group, and were not designed to generalize to all Medicare cases.

The two DERs were the proportions of the same 100 cases found to have diagnostic errors by UPRO review and Iliad review respectively. Because it was likely that the number of diagnostic errors found by one review method in a set of cases would be
correlated to the number of diagnostic errors found by another review method, McNemar's chi-square test for comparing proportions in two paired groups was used to analyze the statistical significance of the difference between the two DERs. The level of significance ($\alpha$) was set at 0.05 for a one-tailed test.\(^{164}\)

The two TPRs were the proportions of cases flagged by nurse review that were confirmed correct by physician review, for UPRO review and Iliad review, respectively. The review processes were independent. Hence, analysis for statistical significance was performed using the z-test for difference between two independent proportions. The level of significance ($\alpha$) was set at 0.05 for a one-tailed test.\(^{164}\)

**Iliad Gold Standard Review**

The gold standard reviewers in this experiment were all volunteer physicians from the University of Utah School of Medicine. There were four specialists in internal medicine: a gastroenterologist, a cardiologist, a pulmonary physician and an endocrinologist. None were on UPRO's panel of physician reviewers. All had teaching responsibilities and involvement in other Iliad research, and thus were not blinded to Iliad's role in the experiment. They were each asked to review a case for a specific diagnostic error, such as a missed diagnosis or an unlikely diagnosis, and to rate the quality problem according to the PRO severity classification. Each case flagged by an Iliad nurse review was referred to one of these four physicians according to specialty.

**Software and Hardware**

Iliad version 2.03, August 1990 was used on a Macintosh SE30 computer. After all 100 cases were entered, the cases were run through a newer version, Iliad version 3.0, dated March 20th, 1991 with the knowledge base dated March 14th, 1991, to take advantage of improvements to the program made while the data entry was in progress. The diagnostic lists generated by this later Iliad for the 100 cases were used to compare to the attending physician's diagnoses.
Experiment II: Comparing Iliad and UCDS Reviews

Case Selection

UPRO was evaluating UCDS as a nurse review tool. UPRO randomly selected 6% of all Utah Medicare hospitalizations in the period of February, 1991 to July, 1991 to undergo the phase I testing of the UCDS. In the UCDS review, the nurse screening was performed by UPRO nurses using the UCDS program, and potential problem cases flagged by the program were referred to UPRO physician reviewers. In the phase I testing, 1,037 cases went through the UCDS review. A proportion of these 1,037 cases would have been flagged by the UCDS nurse review as containing potential problems and referred for physician review, and the remaining cases classified as problem-free. The UCDS physician review would have confirmed problems in some of the referred cases, and denied the presence of problems in the rest. This was similar to the pattern seen in the manual UPRO review, from which cases were used in Experiment I. An important difference was that those cases classified by the UCDS nurse review as problem-free were kept and thus available for use in this second experiment.

Each of the 1,037 cases was looked at to see if it was suitable for Iliad review. This was to provide an indication of the proportion of Medicare cases Iliad could be applied to, since the 1,037 cases were obtained by a 6% random sampling of all Utah Medicare cases in a 6-month period. To be eligible for the experiment, the principal diagnosis in the case had to be one contained in Iliad's knowledge base. Also, the case could not be completely void of diagnostic work-up information (history, physical examination findings, and laboratory test results). According to the above criteria, out of the 1,037 cases, 666 (or 64.2%) were found to be suitable for Iliad review.

Five hundred cases were randomly selected from among these 666 cases containing diagnoses Iliad could recognize. These 500 cases consisted of two groups, those considered by the UCDS nurse review to be problem-free (called non-UCDS-nurse-flagged), and those felt to have management problems by the UCDS nurse review and
referred for physician review (called UCDS-nurse-flagged cases). From the non-UCDS-nurse-flagged cases, a random selection of 50 cases were chosen for Iliad review. It was thought that because diagnostic errors led to quality problems in management, it would be more likely to find diagnostic errors in the UCDS-nurse-flagged cases. If comparing the results of the Iliad review for the UCDS-nurse-flagged cases and the non-UCDS-nurse-flagged cases showed this expectation to be fulfilled, then it might be more cost-effective to use Iliad only in those cases already flagged by the PRO nurse review. To conserve the limited resources available for this experiment, therefore, it was opted to review a smaller number of randomly selected non-UCDS-nurse-flagged cases, but to review all UCDS-nurse-flagged cases.

At the end of the experiment, it was discovered that four of the 50 non-UCDS-nurse-flagged cases were initially mistakenly classified as such by UPRO, and had actually been flagged for physician review by the UPRO nurses. Thus, 46 randomly selected non-UCDS-nurse-flagged cases remained for this second experiment. There were 220 non-UCDS-nurse-flagged cases among the 500 cases.

Of the 500 cases, 280 were flagged by the UCDS nurse review and referred for physician review (UCDS-nurse-flagged cases). All 280 UCDS-nurse-flagged cases were used in the experiment. Note that this group was analogous to the 100 cases used in Experiment I. Therefore, a total of 326 cases were used in Experiment II.

**Sample Size**

As previously calculated for Experiment I (in Figure 4), the minimal sample size required to detect a 1% diagnostic error rate would be 70 cases. However, additional funding had become available for this second experiment, making it possible to pay UPRO physicians to perform the gold standard physician review (for Iliad), instead of using volunteer physicians from the University of Utah as in the first experiment. Feedback from UPRO nurse reviewers had indicated a more conservative attitude toward certifying errors on the part of UPRO physicians, so it was decided to include as many cases in the
experiment as could be afforded. Preliminary experience had indicated that Iliad nurse review would flag approximately a third of cases, and since the budget allowed for 100 physician reviews, 300 cases would be selected for this second experiment. Because the Iliad nurse review took time and effort, and it was believed that diagnostic errors were more likely found in cases already flagged for management problems, the decision was made to obtain the majority of the 300 cases needed from the UCDS-nurse-flagged group. For comparison, a random sample would be drawn from the non-UCDS-nurse-flagged group. Feedback from the UPRO nurses involved in UCDS testing had indicated a 50% referral rate, thus it was expected that sampling 500 cases would yield 250 UCDS-nurse-flagged cases. An additional 50 non-UCDS-nurse-flagged cases would be randomly picked from the rest. The total number of diagnostic errors found by Iliad review would be calculated and projected back to the original sample size of 500.

Research Hypotheses

Experiment II planned to indicate the future usefulness of Iliad review by comparing it to UCDS review, which was targeted for eventual nationwide implementation in PROs. Similar to Experiment I, the effectiveness and efficiency of Iliad review were tested with the following hypotheses:

IIA: Iliad review detected a higher Diagnostic Error Rate (DER₁) for the test population, compared to the Diagnostic Error Rate found by the UCDS review (DER₀).

Null hypothesis (H₀): DER₁ ≤ DER₀
Alternative (experiment) hypothesis (H₁): DER₁ > DER₀

IIB: Iliad nurse review had a higher TPR (TPR₁), compared to the TPR produced by the UCDS nurse review (TPR₀).

Null hypothesis (H₀): TPR₁ ≤ TPR₀
Alternative (experiment) hypothesis (H₁): TPR₁ > TPR₀
**Statistical Analysis**

Similar to Experiment I, the outcomes of this second experiment were the Diagnostic Error Rate found by the Iliad review, compared to that found by the UCDS review; and the TPR of Iliad nurse review, compared to that produced by the UCDS nurse review. The test population was randomly selected from those, in a 6% random sampling of Utah Medicare inpatient cases, with a diagnosis Iliad could recognize. Thus the DERs and TPRs calculated in this experiment applied were generalizable to Utah Medicare inpatient cases with Iliad-recognized diagnoses.

Again, McNemar's chi-square test for comparing proportions in two paired groups was used to analyze the statistical significance of the difference between the two DERs. For the difference between the two TPRs, statistical significance was tested with the z-test for difference between two independent proportions. For both tests, the level of significance ($\alpha$) was set at 0.05 for a one-tailed test.\(^{164}\)

**Iliad Gold Standard Review**

Six UPRO physicians, either general internists or family practitioners, agreed to perform the gold standard review in this experiment. None of them were University of Utah physicians (verbal communication with Dr. A. Lloyd Poulsen, medical director of UPRO). This investigator was not aware of their identity, and they were also not told about Iliad. The gold standard review form sent to these reviewers contained one or more questions, of the form: "The diagnosis of XXX was inadequate in the case. Is this a diagnostic error?" or "The diagnosis of YYY was not made in the case. Is this a diagnosis missed by the attending physician?". These questions were analogous to those asked by UPRO nurses, in the routine PRO Medicare reviews, requiring decisions by the UPRO physicians. Each case flagged by an Iliad nurse review was referred to one of these six physicians through UPRO. This investigator was blinded to the case assignments, which was handled totally by UPRO. The review form was also returned from UPRO with the physician identifier removed.
Software and Hardware

Iliad version 3.1 dated July 12th, 1991, with the knowledge base dated same, was used on a Macintosh SE30 computer. After about half the cases were entered, a Macintosh Quadra 700 computer became available, and the remaining cases were entered in Iliad version 4.0 dated December 19, 1991 with the knowledge base dated December 17, 1991. As a result of both the faster computer and the newer, faster version of Iliad, data entry time was shortened. After all 326 cases were entered, the cases were run through an experimental version of Iliad: version 4.1, dated January 19th, 1993 with the knowledge base dated same. This version of Iliad provided ICD9- (International Classification of Diseases, 9th revision) codes along with diagnoses. The diagnostic lists generated by this later Iliad for the 326 cases were used to compare to the attending physician's diagnoses.

Protection of Human Subjects

Human subjects were not directly involved in this research. However, the experiments examined actual patient charts and the diagnostic performance of real physicians. In accordance with standard UPRO procedures, the identity of the patients and physicians were completely protected. Physicians' and patients' names and other identifiers were not attached to the results. Physicians would not receive UPRO censure as a result of any quality problems detected. The review process was accomplished after the patient care episode was ended, and hence would not influence the patient care in any way.
RESULTS

Experiment I: Comparing Iliad and UPRO Reviews

Patient Demographics

One hundred Medicare cases were used in this experiment. The ages of the 100 patients ranged from 31 to 102 years. The average age was 76. Of the 100 patients, 41 were female. The length of stay ranged from 2 to 27 days, with an average of 8 days. There were five deaths among the 100 cases. The admissions were to 22 (out of a total of 54) hospitals in Utah. There were 31 different principal discharge diagnoses among the 100 cases. The most common principal diagnosis was unstable angina (28 cases), followed by pneumonia (18 cases) and benign prostatic hypertrophy (6 cases).

Results of UPRO and Iliad Reviews

Diagnostic error rates and true positive rates. The test population were 100 cases randomly selected from the Utah Medicare inpatient cases previously flagged by the UPRO nurse review as containing quality problems and referred for physician review, and which contained a diagnosis recognizable by Iliad. The results of the URPO physician review were summarized in Table 4. For this test population, none of the quality problems found by UPRO review was diagnostic in nature.

Iliad review of the 100 cases, using the previously described 80%/20% threshold criteria for discrepancy between Iliad's and the attending physician's diagnoses, flagged 28 cases for the gold standard physician review. Each case was reviewed by a University of Utah physician, matched by subspecialty. Because these were volunteer physicians who were working with the Department of Medical Informatics on various other research projects, it was not possible to blind them to the sources of the diagnostic error (Iliad
Table 4. Experiment I: results of UPRO and Iliad reviews.

<table>
<thead>
<tr>
<th>Review process</th>
<th>UPRO</th>
<th>Iliad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of cases flagged by nurse review</td>
<td>100</td>
<td>28</td>
</tr>
<tr>
<td>Number of cases with appropriate care</td>
<td>72</td>
<td>11</td>
</tr>
<tr>
<td>Number of cases with quality problems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level I (score 1)</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>Level II (score 5)</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Level III (score 25)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>DER of test population</td>
<td>0%</td>
<td>17%</td>
</tr>
<tr>
<td>TPR of nurse review</td>
<td>28.0%</td>
<td>60.7%</td>
</tr>
<tr>
<td>Average severity score of cases with problems</td>
<td>2.0</td>
<td>5.3</td>
</tr>
</tbody>
</table>

versus attending physician). Hence, a reviewer bias could not be excluded. The results of the gold standard physician review were summarized in Table 4. Because only those cases previously flagged by UPRO nurses (and containing a diagnosis Iliad could recognize) were selected for the experiment, it was possible to calculate a Diagnostic Error Rate (DER) for the test population only. This DER might not reflect the actual number of diagnostic errors in the general Medicare population.

The difference in DER of the two reviews, for the test population, was statistically significant ($p < 0.001$). The difference in TPRs was also statistically significant ($p < 0.001$).

**Quality problems found by the reviews.** As expected, the UPRO review and the Iliad review detected quality problems of a different nature. The 28 quality problems found by UPRO review were summarized in Table 5, and the 17 quality problems found by the Iliad review in Table 6. Interestingly, a missed diagnosis by the attending physician appeared to result in more serious quality problems than if an unlikely diagnosis was made (see Table 7). The mean severity of the quality problems resulting from a missed diagnosis
Table 5. Experiment I: quality problems found in the UPRO review.

Level I problem: 21 cases

- Inadequate discharge planning: 3
- Medical stability not assured at discharge: 9
- Medication error: 1
- Failure to perform a test: 1
- Inadequate documentation: 5
- Failure to perform a physical examination: 1
- Nosocomial infection: 1

Level II problem: 7 cases

- Medical stability not assured at discharge: 4
- Failure to perform a test: 2
- Medication error: 1

Table 6. Experiment I: quality problems found in the Iliad review.

Level I problem: 10 cases

- Missed acute myocardial infarction (AMI): 1
- Missed osteoarthritis: 1
- Missed severe secondary hyperparathyroidism: 1
- Unlikely diagnosis of chronic lymphocytic leukemia: 1
- Unlikely diagnosis of acute gastritis: 1
- Unlikely diagnosis of peritonitis: 1
- Unlikely diagnosis of AMI: 2
- Unlikely diagnosis of diverticulitis: 2

Level II problem: 5 cases

- Missed AMI: 1
- Missed diabetes mellitus: 2
- Missed diabetic nephropathy: 1
- Missed severe secondary hyperparathyroidism: 1

Level III problem: 2 cases

- Missed pulmonary embolus: 1
- Missed lung cancer: 1
Table 7. Experiment I: severity of missed diagnoses and unlikely diagnoses.

<table>
<thead>
<tr>
<th>Quality problem</th>
<th>Level I (score 1)</th>
<th>Level II (score 5)</th>
<th>Level III (score 25)</th>
<th>Total number</th>
<th>Average score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missed diagnosis</td>
<td>3 (30%)</td>
<td>5 (50%)</td>
<td>2 (20%)</td>
<td>10 (100%)</td>
<td>7.8</td>
</tr>
<tr>
<td>Unlikely diagnosis</td>
<td>7 (100%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>7 (100%)</td>
<td>1.0</td>
</tr>
</tbody>
</table>

was 7.8, whereas that resulting from an unlikely diagnosis was 1.

Comparing the two reviews. The results of the UPRO review and the Iliad review were compared to see if the same quality problems were detected by the two independent processes (see Table 8). Six cases were found to have quality problems detected by both review procedures.

Among the six cases with quality problems detected by both the UPRO and Iliad review, there was none for which the same quality problem was detected by the two procedures. That is, UPRO review and Iliad review detected quality problems of different nature (see Table 9). For instance, in one of these six cases, Iliad review detected an acute

Table 8. Experiment I: comparing the results of Iliad and UPRO reviews.

<table>
<thead>
<tr>
<th>Iliad review:</th>
<th>Problem detected</th>
<th>Problem not detected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPRO review:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem detected</td>
<td>6</td>
<td>22</td>
<td>28</td>
</tr>
<tr>
<td>Problem not detected</td>
<td>11</td>
<td>61</td>
<td>72</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>83</td>
<td>100</td>
</tr>
<tr>
<td>Quality problem detected by Iliad</td>
<td>Quality problem detected by UPRO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missed: AMI (Level I)</td>
<td>Consent for procedure not in record (Level I)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missed: lung cancer (Level III)</td>
<td>Failure to perform an investigation (Level II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missed: hyperparathyroidism (Level II)</td>
<td>Stability at discharge not assured (Level I)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unlikely: AMI (Level I)</td>
<td>Stability at discharge not assured (Level II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unlikely: diverticulitis (Level I)</td>
<td>Stability at discharge not assured (Level I)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missed: AMI (Level I)</td>
<td>Consent for procedure not in record (Level I)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

myocardial infarction (AMI) which the attending physician missed. This physician had diagnosed unstable angina and had then proceeded to a cardiac catheterization, coronary angiogram and angioplasty. The UPRO nurse reviewer had flagged the case because the treating physician failed to obtain consent for angioplasty. The physician reviewer agreed with the nurse's opinion and did not have any additional comments. The case was referred for a gold standard physician review by a cardiologist from the University of Utah School of Medicine, who agreed with Iliad's diagnosis of AMI because the cardiac enzymes and electrocardiogram clearly showed evidence of infarction. The gold standard reviewer noted that the cardiac catheterization might not have been warranted in view of the increased risk in AMI. The original UPRO reviewer had also overlooked the AMI, perhaps because his attention was focused by the nurse reviewer's question regarding informed consent.

Since there was no overlap between the types of quality problems detected by the UPRO and the Iliad reviews, the total number of quality problems detected by using both procedures would be increased to \((6 + 11 + 22)\) or 39, out of the 100 cases reviewed (see Tables 8 and 10). McNemar's chi-square test showed that the increase in the proportion of cases found to have quality problems by using the combined review, compared to the
Table 10. Experiment I: combined UPRO and Iliad review and cost estimation.

<table>
<thead>
<tr>
<th>Review process</th>
<th>UPRO</th>
<th>Iliad</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of quality problems found</td>
<td>28</td>
<td>17</td>
<td>39</td>
</tr>
<tr>
<td>Average severity score of problem cases</td>
<td>2.0</td>
<td>5.3</td>
<td>3.4</td>
</tr>
<tr>
<td>Level I problems (score 1)</td>
<td>21</td>
<td>10</td>
<td>26</td>
</tr>
<tr>
<td>Level II problems (score 5)</td>
<td>7</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>Level III problems (score 25)</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Average nurse review time (min.)</td>
<td>20.0</td>
<td>29.2</td>
<td>49.2</td>
</tr>
<tr>
<td>Average physician review time (min.)</td>
<td>36.7</td>
<td>8.8</td>
<td>45.5</td>
</tr>
<tr>
<td>Cost of nurse review ($15/hr.)</td>
<td>$500.00</td>
<td>$730.00</td>
<td>$1230.00</td>
</tr>
<tr>
<td>(100 cases)</td>
<td>(100 cases)</td>
<td>(100 cases)</td>
<td></td>
</tr>
<tr>
<td>Cost of physician review ($25/case)</td>
<td>$2500.00</td>
<td>$700.00</td>
<td>$2500.00</td>
</tr>
<tr>
<td>(100 cases)</td>
<td>(28 cases)</td>
<td>(100 cases)</td>
<td></td>
</tr>
<tr>
<td>Total cost</td>
<td>$3000.00</td>
<td>$1430.00</td>
<td>$3730.00</td>
</tr>
<tr>
<td>Number of problem cases confirmed</td>
<td>28</td>
<td>17</td>
<td>39</td>
</tr>
<tr>
<td>Cost per confirmed problem case</td>
<td>$107.14</td>
<td>$84.11</td>
<td>$95.64</td>
</tr>
</tbody>
</table>

proportion found by the UPRO review alone, was also statistically significant (p < 0.01).

In addition to increasing the number of quality problems found, using Iliad in addition to the PRO review would result in more serious quality errors being found (see Table 10).

Cost estimation. In the UPRO review, the nurses required 5 to 40 minutes per review, with an average of 20 minutes (verbal communication with UPRO nurse reviewers). The physician review averaged 36.7 minutes, ranging from 15 to 153 minutes. For the Iliad review, 15 to 40 minutes were needed to enter the data from one patient record into Iliad, with an average of 29.2 minutes. The gold standard physician review required 2 to 20 minutes per case, averaging 8.8 minutes. An estimation of the cost incurred for each review and for a combined review was presented in Table 10. The cost-benefit estimation
in this experiment was for using Iliad only in cases already flagged by UPRO nurses for physician review. The cost of nurse review was estimated at $15.00 an hour, and that of physician review $25.00 per case.

Although using Iliad in addition to standard PRO review resulted in an increase in net cost, the cost per quality problem confirmed decreased. McNemar’s chi-square test showed the increase in proportion of cases with quality problems found by the combined review over the UPRO review alone to be statistically significant as well (p < 0.01).

**Interrater Reliability in Matching Diagnoses**

The Iliad nurse review was performed by this author. After all 100 cases were entered into Iliad, Iliad’s diagnoses and the corresponding attending physician’s diagnoses were tabulated for comparison according to the 80%/20% threshold criteria explained earlier. Because of the large variety of medical terminology, the diagnostic terms in Iliad could be expected to differ slightly from those used by the attending physicians. Thus, the two sets of diagnoses could not be compared merely by exact matching of terminology, but required some interpretation by the Iliad nurse reviewer. For instance, "ischemic cardiomyopathy" in Iliad was matched to "coronary atherosclerosis" in the chart, instead of flagging the case for missing the diagnosis of ischemic cardiomyopathy, rated more than 80% probability in Iliad; and for making the unlikely diagnosis of coronary atherosclerosis, rated less than 20% probability (in fact, not present) in Iliad.

To see if another reviewer might arrive at the same flagging decisions, the list comparing the attending physician’s diagnoses with Iliad’s diagnoses of the 100 cases was given to a general practitioner. His instructions were to pick the cases violating the 80%/20% threshold criteria. No other suggestions for matching, such as the example given above, were provided. The two set of flagging decisions were compared. There was agreement in 76% of the flagging decisions. The reliability of the two rater’s judgments was determined using Cohen’s kappa. This statistic measured interrater reliability while correcting for chance level of agreement between two independent
judgments. A kappa of 0 would indicate that agreement was by chance only, whereas a kappa of 1 would indicate complete agreement without chance operating, and -1 would indicate complete disagreement without chance operating.\textsuperscript{166,167} Cohen's kappa was calculated to be 0.38, which was statistically significant (p < 0.001 for a one-tailed test).

**Experiment II: Comparing Iliad and UCDS Reviews**

**Patient Demographics**

Three hundred and twenty six Medicare cases were used in this experiment. The ages of the 326 patients ranged from 34 to 103 years. The average age was 75. One hundred and sixty five of the 326 patients were female (50.6%). The length of stay ranged from one to 16 days, with an average of 5.2 days. The admissions were to 36 (out of 54) hospitals in Utah. The 10 most common principal diagnoses are listed in Table 11.

**Results of UCDS and Iliad Reviews**

**Generalization of Iliad review results.** The test population were 500 cases randomly selected from those, of a random sampling of Utah Medicare inpatient cases in a 12-month period, containing a diagnosis recognizable by Iliad. These 500 cases were all sent through the UCDS nurse review, and 280 were flagged for UCDS physician review.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Diagnosis</th>
<th>Number of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Congestive heart failure</td>
<td>34</td>
</tr>
<tr>
<td>2</td>
<td>Pneumonia</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Cerebrovascular accident / transient ischemic attack</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>Angina</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>Acute myocardial infarction (AMI)</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>Cardiac dysrhythmia</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>Malignancy</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>Peptic ulcer disease / gastrointestinal hemorrhage</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>Chest pain</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>Cholecystitis</td>
<td>9</td>
</tr>
</tbody>
</table>
The results of UCDS review were summarized in Table 12. For this test population, four of the quality problems found by UCDS review were diagnostic in nature, hence giving a Diagnostic Error Rate (DER) of 4/500 or 0.8%.

All 280 UCDS-nurse-flagged cases were also sent through the Iliad review, and 46 were randomly selected from the remaining 220 non-UCDS-nurse-flagged cases for Iliad review. Thus, 326 cases went through the Iliad nurse review. Using the previously described 80%/20% threshold criteria for discrepancy between Iliad's and attending physician's diagnoses, 85 cases were flagged in the UCDS-nurse-flagged group and 3 in the non-UCDS-nurse-flagged group.

Each case flagged by the Iliad nurse review was referred for a gold standard physician review, performed by a UPRO physician recruited for the experiment. Six UPRO physicians participated in the gold standard review. They were all general

Table 12. Experiment II: Results of UCDS and Iliad reviews.

<table>
<thead>
<tr>
<th>Review process</th>
<th>UCDS</th>
<th>Iliad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cases sent through nurse review</td>
<td>500</td>
<td>326</td>
</tr>
<tr>
<td>Number of cases flagged by nurse review</td>
<td>280</td>
<td>88</td>
</tr>
<tr>
<td>Number of cases with appropriate care</td>
<td>269</td>
<td>65</td>
</tr>
<tr>
<td>Number of cases with quality problems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level I (score 1)</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>Level II (score 5)</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Level III (score 25)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of diagnostic errors</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>Average severity score of cases with problems</td>
<td>4.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Average severity score of diagnostic errors</td>
<td>5.0</td>
<td>1.5</td>
</tr>
<tr>
<td>DER of cases reviewed</td>
<td>0.8%</td>
<td>7.1%</td>
</tr>
<tr>
<td>DER projected for test population</td>
<td>0.8%</td>
<td>6.0%</td>
</tr>
<tr>
<td>TPR of cases reviewed</td>
<td>3.9%</td>
<td>26.1%</td>
</tr>
<tr>
<td>TPR of nurse review projected for test population</td>
<td>3.9%</td>
<td>30.6%</td>
</tr>
</tbody>
</table>
practitioners with no teaching appointments at the University of Utah School of Medicine. The UPRO physicians were asked to review cases for diagnostic errors and to rate the quality problem according to the PRO severity classification. However, they were not told about Iliad, in an effort to avoid the positive reviewer bias that might have been present in Experiment I. The gold standard physician review results are also summarized in Table 12.

The test population was the 500 cases that were randomly selected from a random sampling of Utah Medicare inpatient cases during a 6-month period, and which contain diagnoses recognizable by Iliad. Therefore, the results of Experiment II were generalizable to all Medicare inpatient cases with a diagnosis recognized by Iliad. The DER for the test population was calculated in Table 13. The difference between the DER found by Iliad review in the UCDS-nurse-flagged group and the DER in the non-UCDS-nurse-flagged

Table 13. Experiment II: generalization of Iliad review results to test population.

<table>
<thead>
<tr>
<th>Total sample size for Iliad review = 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sample size for Iliad review = 500</td>
</tr>
<tr>
<td>UCDS-nurse-flagged = 280</td>
</tr>
<tr>
<td>Non-UCDS-nurse-flagged = 220</td>
</tr>
<tr>
<td>Sampled for Iliad nurse review = 280</td>
</tr>
<tr>
<td>Sampled for Iliad nurse review = 46</td>
</tr>
<tr>
<td>Iliad nurse review flags = 85</td>
</tr>
<tr>
<td>Iliad nurse review flags = 3</td>
</tr>
<tr>
<td>Problems = 21 (18 Level I, 3 Level II)</td>
</tr>
<tr>
<td>Problem cases found = 2 (both Level I)</td>
</tr>
<tr>
<td>TPR for this group = 21/85 = 24.7%</td>
</tr>
<tr>
<td>TPR for this group = 2/3 = 66.7%</td>
</tr>
<tr>
<td>DER for this group = 21/280 = 7.5%</td>
</tr>
<tr>
<td>DER for this group = 2/46 = 4.3%</td>
</tr>
<tr>
<td>Problems in whole group = 21</td>
</tr>
<tr>
<td>Projected problems in group = 4.3% X 220 = 9</td>
</tr>
<tr>
<td>Projected flags for whole group = 9/66.7% = 13</td>
</tr>
</tbody>
</table>

Number of problems projected for entire sample = 21 + 9 = 30
DER for test population = 30/500 = 6%
Projected TPR for test population = 30/(85+13) = 30.6%
group was found to be statistically not significant, using the z-test for difference between two independent proportions. Power was found to be 0.99.\(^{164}\)

**Diagnostic error rates and true positive rates.** Because the test population was randomly sampled, the results of Experiment II were generalizable to all Medicare inpatient cases with a diagnosis recognized by Iliad. The difference between the projected DER for the test population, found by Iliad review, and the DER found by UCDS review was statistically significant \((p < 0.001)\). The difference in TPRs was also statistically significant \((p < 0.001)\). However, the Iliad review appeared to detect less serious problems than those found by the UCDS review.

**Quality problems found by the reviews.** The quality problems found by UCDS review are summarized in Table 14. Unfortunately, two of the UCDS physician review worksheets were not available, thus details of two problems were not known. Interestingly, four diagnostic errors were found among the 11 quality problems detected by UCDS review. The questions raised in the UCDS nurse review and the physician reviewer's responses are summarized in Table 15. The 23 quality problems found by the Iliad review are presented in Table 16.

Again, a missed diagnosis by the attending physician appeared to result in more serious quality problems than if an unlikely diagnosis was made (see Table 17).

Table 14. Experiment II: quality problems found in the UCDS review.

<table>
<thead>
<tr>
<th>Level I problem: 1 case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure to perform a test: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level II problem: 10 cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure to diagnose pulmonary embolus: 1</td>
</tr>
<tr>
<td>Failure to diagnose gouty arthritis: 1</td>
</tr>
<tr>
<td>Failure to diagnose azotemia: 1</td>
</tr>
<tr>
<td>Failure to diagnose pulmonary infection/fluid overload: 1</td>
</tr>
<tr>
<td>Failure to perform a test: 3</td>
</tr>
<tr>
<td>Medical stability not assured at discharge: 1</td>
</tr>
<tr>
<td>(Details not known: 2)</td>
</tr>
</tbody>
</table>
Table 15. Experiment II: diagnostic errors found in the UCDS review.

<table>
<thead>
<tr>
<th>Nurse review concern</th>
<th>Physician review judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical stability not assured at discharge.</td>
<td>No problem with stability at discharge, but diagnosis of pulmonary embolus missed.</td>
</tr>
<tr>
<td>Antibiotics started without culture.</td>
<td>Agreed, and diagnosis of gouty arthritis missed.</td>
</tr>
<tr>
<td>Chest x-ray indicated but not done.</td>
<td>Agreed, and diagnosis of pulmonary fluid overload missed.</td>
</tr>
<tr>
<td>Premature discharge with raised blood urea and creatinine levels.</td>
<td>Agreed, and diagnosis of azotemia missed.</td>
</tr>
</tbody>
</table>

Table 16. Experiment II: quality problems found in the Iliad review.

Level I problem: 20 cases

- Failure to diagnose hypertensive heart disease: 1
- Failure to diagnose sporadic hypertriglyceridemia: 2
- Failure to diagnose multi-infarct dementia: 1
- Failure to diagnose prerenal azotemia: 1
- Failure to diagnose urinary tract infection (UTI): 1
- Unlikely diagnosis of asthma: 1
- Unlikely diagnosis of congestive heart failure: 5
- Unlikely diagnosis of diabetes mellitus: 1
- Unlikely diagnosis of renal failure: 2
- Unlikely diagnosis of bilateral pneumonitis: 1
- Unlikely diagnosis of angina pectoris: 1
- Unlikely diagnosis of UTI: 2
- Unlikely diagnosis of acute myocardial infarction (AMI): 1

Level II problem: 3 cases

- Failure to diagnose AMI: 1
- Failure to diagnose acquired hyperlipidemia: 1
- Unlikely diagnosis of gastritis / duodenitis: 1
Table 17. Experiment II: severity of missed diagnoses and unlikely diagnoses.

<table>
<thead>
<tr>
<th>Quality problem</th>
<th>Level I (score 1)</th>
<th>Level II (score 5)</th>
<th>Level III (score 25)</th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Missed diagnosis</td>
<td>6 (71.4%)</td>
<td>2 (28.6%)</td>
<td>0 (0%)</td>
<td>8</td>
<td>(100%)</td>
</tr>
<tr>
<td>Unlikely diagnosis</td>
<td>14 (92.9%)</td>
<td>1 (7.1%)</td>
<td>0 (0%)</td>
<td>15</td>
<td>(100%)</td>
</tr>
</tbody>
</table>

Comparing the two reviews. The results of the UCDS review and the Iliad review were compared to see if the same quality problems were detected by the two independent processes. None of the cases was found to have problems by both reviews. Thus, the Iliad review did not detect the four diagnostic problems found by UCDS review.

Cost estimation. In the UCDS review, the nurses required 20 to 180 minutes per review, with an average of 100 minutes. The physician review averaged 30 minutes, ranging from 10 to 45 minutes (verbal communication with UPRO staff involved in the UCDS phase I testing project). For the Iliad review, 9 to 77 minutes were needed to enter the data from one patient record into Iliad, with an average of 24.3 minutes. The gold standard physician review required 1 to 25 minutes per case, averaging 13 minutes. An estimation of the cost incurred for each review and for a combined review is presented in Table 18. The cost-benefit estimation for the combined review was calculated twice, first for only those cases already flagged by UCDS nurse review, then projected for all 500 cases (see Table 18). The cost of nurse review was again estimated at $15.00 an hour, and that of physician review $25.00 per case. Although using Iliad in addition to UCDS review would result in an increase in net cost, the cost per problem case confirmed would decrease. This would be true whether Iliad review was applied to only those cases already flagged by UCDS nurse review, or to all cases sent through the UCDS nurse review.
Table 18. Experiment II: combined UCDS and Iliad review and cost estimation.

<table>
<thead>
<tr>
<th>Review process</th>
<th>UCDS</th>
<th>UCDS with Iliad review for all cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of problem cases found</td>
<td>11</td>
<td>31 (11+20)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40 (11+29)</td>
</tr>
<tr>
<td>Average severity score of problem cases</td>
<td>4.6</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.3</td>
</tr>
<tr>
<td>Number of Level I problems</td>
<td>1</td>
<td>18 (1+17)</td>
</tr>
<tr>
<td>Number of Level II problems</td>
<td>10</td>
<td>13 (10+3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13 (10+3)</td>
</tr>
<tr>
<td>Average nurse review time (min.)</td>
<td>100</td>
<td>113.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>124.3</td>
</tr>
<tr>
<td>Average physician review time (min.)</td>
<td>30</td>
<td>33.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32.9</td>
</tr>
<tr>
<td>Cost of nurse review ($15/hr.)</td>
<td>$12500.00</td>
<td>$14200.00</td>
</tr>
<tr>
<td></td>
<td>(500 cases)</td>
<td>(500 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$15537.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(500 cases)</td>
</tr>
<tr>
<td>Cost of physician review ($25/case)</td>
<td>$7000.00</td>
<td>$7000.00</td>
</tr>
<tr>
<td></td>
<td>(280 cases)</td>
<td>(280 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$7325.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(293 cases)</td>
</tr>
<tr>
<td>Total cost</td>
<td>$19500.00</td>
<td>$21200.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$22862.50</td>
</tr>
<tr>
<td>Cost per confirmed problem case</td>
<td>$1772.72</td>
<td>$683.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$571.56</td>
</tr>
</tbody>
</table>

**Matching Diagnoses by ICD9 Codes**

In Experiment I, an estimate of the subjectiveness of manually matching Iliad's diagnoses to those of the attending physicians was made by comparing the matches made by the author with those by another general practitioner. In Experiment II, it was decided to see if a more objective way of diagnosis matching could be found. Since almost all Medicare charts ICD9-coded the discharge diagnoses, an experimental version of Iliad which supplied ICD9 codes for its diagnoses was used in Experiment II, so that diagnoses could be compared by ICD9 codes instead of terminology.\(^{165}\)

When matching diagnoses solely by ICD9 codes was tried on the 280 UCDS-nurse-flagged cases, an additional 60 cases were unnecessarily flagged (false positives), apart from the original 85 flags. This was largely because there were mistakes in coding in the charts, for instance, when the chart diagnosis was non-insulin-dependent diabetes
mellitus, but coded as insulin-dependent diabetes mellitus. Also, some ICD9 codes were missing from the charts, as well as in Iliad. Thus, matching solely by ICD9 codes was unlikely to be universally applicable or cost-effective. On the other hand, frequently, when the chart diagnosis was vague or when terminology differed between Iliad's diagnoses and the attending physician's, identical ICD9 codes helped matched the diagnoses. An example was when the chart diagnosis of "severe diarrhea - Clostridia and chemotherapy induced" was matched by ICD9 code to "pseudomembranous colitis" in Iliad. Thus, ICD9 codes were useful as a matching aid, but not as the sole matching criterion.

Varying the Thresholds for Error Flagging

In the Iliad nurse review, the "80%/20%" threshold criteria used to flag cases for physician review referred to the following criteria:

- A diagnosis on the attending physician's list did not reach above 20% probability on Iliad's list - flagging for unlikely diagnosis made by the attending physician.
- A diagnosis exceeding 80% probability on Iliad's list did not appear on the attending physician's list - flagging for a diagnosis missed by the attending physician.

These criteria were empirically chosen, based on past experience with Iliad. As the thresholds were tightened by raising the threshold for flagging missed diagnoses (80% to 85%, 90%, 95%), and by lowering the threshold for flagging unlikely diagnoses (20% to 15%, 10%, 5%), the number of false positive flags would be expected to decrease, at the cost of not detecting some real problems. The effect of tightening the thresholds on DER (Diagnostic Error Rate) and nurse review TPR (True Positive Rate), calculated for the 326 cases reviewed by Iliad, is shown in Table 19 and Figure 5. The DER ranged from 7.1% to 5.5%. The TPR ranged from 27.1% to 24.7%. The arrows in Figure 5 point to the thresholds that resulted in the highest DER and TPR. When 80% was used as the threshold for flagging missed diagnoses, and 15% as the threshold for flagging unlikely
Table 19. Effect of different threshold criteria on DER and TPR of Iliad review.

<table>
<thead>
<tr>
<th>Thresholds</th>
<th>DER(%)</th>
<th>TPR(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80%/20%</td>
<td>7.1</td>
<td>26.1</td>
</tr>
<tr>
<td>80%/15%</td>
<td>7.1</td>
<td>27.1</td>
</tr>
<tr>
<td>85%/20%</td>
<td>6.7</td>
<td>26.2</td>
</tr>
<tr>
<td>80%/10%</td>
<td>6.7</td>
<td>27.1</td>
</tr>
<tr>
<td>85%/15%</td>
<td>6.4</td>
<td>25.6</td>
</tr>
<tr>
<td>90%/20%</td>
<td>6.4</td>
<td>25.6</td>
</tr>
<tr>
<td>85%/10%</td>
<td>6.4</td>
<td>26.3</td>
</tr>
<tr>
<td>80%/5%</td>
<td>6.4</td>
<td>25.3</td>
</tr>
<tr>
<td>95%/20%</td>
<td>6.1</td>
<td>27.0</td>
</tr>
<tr>
<td>90%/15%</td>
<td>6.1</td>
<td>25.0</td>
</tr>
<tr>
<td>90%/10%</td>
<td>6.1</td>
<td>25.6</td>
</tr>
<tr>
<td>85%/5%</td>
<td>6.1</td>
<td>25.3</td>
</tr>
<tr>
<td>95%/15%</td>
<td>5.8</td>
<td>26.4</td>
</tr>
<tr>
<td>95%/10%</td>
<td>5.8</td>
<td>27.1</td>
</tr>
<tr>
<td>90%/5%</td>
<td>5.8</td>
<td>24.7</td>
</tr>
<tr>
<td>95%/5%</td>
<td>5.5</td>
<td>26.5</td>
</tr>
</tbody>
</table>

Figure 5. Iliad review DER and TPR for different flagging thresholds.
diagnoses, the highest DER (7.1%) and TPR (27.1%) resulted. The thresholds chosen for Experiment II, 80%/20%, resulted in the same high DER (the only other threshold combination to do so), but a lower TPR of 26.1%.

Because of financial constraints, loosening the thresholds were not explored, as more false positives would be expected, hence driving up the total cost of the gold standard physician review.

**Interrater Reliability**

**Data entry into Iliad.** The Iliad nurse review could be divided into two parts. First was the data entry of each case into Iliad. Next was the comparing of Iliad's diagnoses to the attending physician's. Experiment I explored the agreement between two independent reviewers in matching diagnoses. When funding became available for a second experiment, it was decided to explore the agreement between reviewers in the data entry into Iliad.

Because of their responsibilities in the PRO Medicare review, it appeared logical to see if UPRO nurses would be able to use Iliad review. Two UPRO nurses entered data from some of the cases used in this experiment into Iliad, after each was given an hour of training on Iliad by the author. In an attempt to avoid introducing bias, the nurses had no further contact with the author during their case entry. The Iliad diagnoses were then compared to the attending physician's by the author, and the cases that violated the 80%/20% threshold criteria were flagged. Each set of flagging decisions, based on the nurse's case entry, was next compared to the author's flagging decisions. A comparison was also made between the flagging decisions of the two UPRO nurses.

Interrater reliability between the UPRO nurses and between each UPRO nurse and the author was calculated using Cohen's kappa. Cohen's kappa was -0.45 (a negative kappa would indicate disagreement) between nurse A and the author, with 31% agreement for 13 cases, statistically not significant at \( \alpha \) of 0.05 for a one-tailed test. Cohen's kappa was -0.33 between nurse B and the author (44% agreement for 16 cases), also not statistically significant at \( \alpha \) of 0.05 for a one-tailed test. Thus the disagreements
were by chance alone. Cohen's kappa between the two nurses was 0.75 (88% agreement for eight cases), statistically significant (p < 0.025, one-tailed test). These results are shown in Figure 6. The major reason for the lack of agreement between the UPRO nurses and the author appeared to be a failure on the part of the UPRO nurses to enter laboratory test and procedure results. For instance, cardiac enzyme levels and electrocardiogram findings were not entered in a patient with acute myocardial infarction.

**Gold standard physician review.** The gold standard physician review was performed with one UPRO physician per case. The six UPRO physicians in Experiment II were not aware of Iliad’s role in the experiment. A sample of cases were independently reviewed by a unique UPRO physician, not one of the six, who was involved in other Iliad research projects and thus knew Iliad. Also, this unique UPRO physician was a pulmonary specialist (but not the one in Experiment I) with teaching responsibilities, whereas the other six UPRO physicians were either general internists or family practitioners. The unique UPRO physician was blinded to the gold standard review results of the other UPRO physicians. Interrater reliability was calculated using Cohen's kappa.\(^{166,167}\) There was 65% agreement on 17 cases, with a kappa of 0.37, statistically significant (at p < 0.025 for a one-tailed test).

In addition, cases flagged by the Iliad nurse review in this second experiment were each independently reviewed by one of the four physicians, at the University of Utah School of Medicine, who served as gold standard reviewers in Experiment I. These university physicians were blinded to the Iliad gold standard review results of the UPRO

![Kappa Diagram]

**Figure 6.** Interrater reliability between different nurse reviewers in data entry into Iliad.
physicians. Interrater reliability between the university physicians and the UPRO physicians for Iliad review was calculated using Cohen's kappa.\textsuperscript{166,167} There was 80\% agreement on 15 cases, with a kappa of 0.44, statistically not significant at $\alpha$ of 0.05 for a one-tailed test.

The unique UPRO physician introduced above (the specialist involved in teaching and in Iliad research) also reviewed seven of the cases reviewed by the university physicians. He was also blinded to the Iliad gold standard review results of the university physicians. Interrater reliability between the university physicians and this unique UPRO physician for Iliad review was calculated using Cohen's kappa.\textsuperscript{166,167} There was 86\% agreement with a kappa of 0.72, statistically significant at $\alpha$ of 0.005 for a one-tailed test.

These explorations of interrater reliability are summarized in Figure 7.

```
Review physician:        UPRO     "Unique"     University of Utah
Kappa:                  0.37     0.72     0.44
                        (significant) (significant) (not significant)
```

Figure 7. Interrater reliability between different gold standard physician reviewers.
DISCUSSION

Research Hypothesis

To summarize, the hypotheses tested in the two experiments were:

Experiment I: IA: Iliad review detected a higher Diagnostic Error Rate (DER) compared to that found by the UPRO review.
IB: Iliad nurse review had a higher True Positive Rate (TPR) compared to that produced by the UPRO nurse review.

Experiment II: IIA: Iliad review detected a higher DER compared to that found by the UCDS review.
IIB: Iliad nurse review had a higher TPR compared to that produced by the UCDS nurse review.

The results of the two experiments supported the hypotheses in each. The results of Experiment I were only applicable to Medicare inpatient cases flagged by UPRO nurse review and suitable for Iliad review (containing a diagnosis recognized by Iliad, and containing diagnostic work-up information). The results of Experiment II were generalizable to all Medicare cases suitable for Iliad review. The results of the two experiments are discussed in the following sections.

Diagnostic Errors Found by Iliad Review

Iliad Review Compared to UPRO and UCDS Reviews

Number of diagnostic errors. It was not surprising that the UPRO and UCDS reviews were unable to detect the diagnostic errors found by Iliad. The PRO review, as performed by UPRO, was not designed with diagnostic errors in mind, but focused on problems of treatment, documentation and discharge (see Tables 2 and 4). The UCDS
review was merely the computerized form of PRO review, and so was expected to perform at the same level of detection for diagnostic errors (see Table 12). On the other hand, Iliad review focused solely on diagnostic errors. It was interesting that the UCDS review did find some diagnostic errors, compared to the earlier UPRO review, which did not find any. The larger number of cases selected from the UCDS review might have been a factor (280 nurse reviewed cases versus 100 cases in the UPRO review). Another reason for the better diagnostic error detection by the UCDS review might have been the fact that the UCDS physician reviewers were aware of taking part in an experiment, and might have been more aggressive in pointing out problems. Nevertheless, Iliad review performed better than either UPRO or UCDS reviews.

**Severity of diagnostic errors.** The diagnostic errors found by Iliad review in the first experiment appeared to have resulted in more serious quality problems, compared to the problems found by UPRO review (see Table 4). However, the diagnostic errors found by Iliad review in the second experiment appeared to have resulted in less serious quality problems, compared to the problems found by UCDS review (see Table 12). The major reason for this reversal was most likely the different Iliad gold standard review physicians used in the two experiments. The four gold standard reviewers in Experiment I were all University of Utah physicians and specialists (a gastroenterologist, a cardiologist, a pulmonary physician and an endocrinologist) involved in teaching. Thus, they might have taken diagnostic errors more seriously, as a major part of medical school training focuses on diagnostic skills. On the other hand, the six UPRO physician used as gold standard reviewers in Experiment II were all general internists or family practitioners experienced in performing PRO review. None of these UPRO physicians were involved in teaching at the University of Utah School of Medicine. Being used to the PRO review, which virtually ignored diagnostic errors to focus on management problems, they might not have thought of diagnostic errors as serious.
A contributing factor to the more severe rating of diagnostic errors found by UCDS review in Experiment II might have been that those diagnostic errors were all found spontaneously by the physician reviewer, and not as a result of a specific nurse question on diagnosis (see Table 15). Perhaps that allowed the physician review to relate the diagnostic error to patient management and outcome, and thus lent more weight to the diagnostic error.

Comparing the Iliad Reviews in the Two Experiments

Number of diagnostic errors. The Iliad review results differed considerably between the two experiments (see Table 20). Although the two experiments also differed in test population and sample size, it was believed that the most significant factor contributing to the difference in Iliad performance was the gold standard physician review. The DER found by Iliad review, as well as the TPR of the Iliad nurse review, depended largely on the judgment of the gold standard physician reviewers. The differences in Iliad review DER and TPR between the two experiments were both statistically significant (at p < 0.003 for a two-tailed test) when tested with the z-test for two independent proportions.\textsuperscript{164}

Table 20. Comparing the Iliad reviews in the two experiments

<table>
<thead>
<tr>
<th>Experiment:</th>
<th>I: compared with UPRO</th>
<th>II: compared with UCDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of cases reviewed</td>
<td>100</td>
<td>326</td>
</tr>
<tr>
<td>Total number of cases flagged by nurse review</td>
<td>28</td>
<td>88</td>
</tr>
<tr>
<td>Number of cases with appropriate care</td>
<td>11</td>
<td>65</td>
</tr>
<tr>
<td>Number of cases with quality problems</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td>Level I (score 1)</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Level II (score 5)</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Level III (score 25)</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>DER of reviewed cases</td>
<td>17.0%</td>
<td>7.1%</td>
</tr>
<tr>
<td>TPR of nurse review</td>
<td>60.7%</td>
<td>26.1%</td>
</tr>
<tr>
<td>Average severity score of cases with problems</td>
<td>5.3</td>
<td>1.5</td>
</tr>
</tbody>
</table>
There were a few reasons that could have contributed to the much better performance of Iliad review, as judged by the gold standard review, in the first experiment. The four gold standard reviewers in Experiment I were all volunteer University of Utah physicians familiar with Iliad research. Hence, they could have had a positive bias for Iliad, as it was not possible to blind them to Iliad's role in the experiment. In addition, these four physicians were specialists (a gastroenterologist, a cardiologist, a pulmonary physician and an endocrinologist) with teaching responsibilities in the University of Utah School of Medicine. None of these University physicians were on UPRO's panel of review physicians. They thus might have a mindset that took diagnostic errors more seriously and be less inclined to accept a faulty diagnostic process.

On the other hand, the six UPRO physician used as gold standard reviewers in Experiment II were all general internists or family practitioners experienced in performing PRO review. In their gold standard review for the experiment, these physicians were not aware that the questioned diagnoses were suggested by Iliad. There might have been a negative bias against this diagnostic review as yet another PRO regulatory activity to complicate the work of physicians. Also, none of these UPRO physicians were teaching at the University of Utah. They might not have thought of diagnostic errors as worthy of notice, as long as the patient management and outcome were adequate - the real-world view. This was supported by the discovery that 17 out of the 20 Level I problems were marked "diagnostic error present" and the "care appropriate", an apparent contradiction. In the PRO severity classification, a Level I problem was one with "no potential adverse effect on the patient", such as a lack of documentation. When a quality problem was present, the least severity rating should be Level I. The category of "care appropriate" was meant to be used for cases with no diagnostic errors, or when the review physician was undecided about the presence of a diagnostic error. In the calculation of DER and TPR for Experiment II, a gold standard review marked "diagnostic error present" and the "care appropriate" was counted as a Level I problem. A review marked "diagnostic error
unclear" with "care appropriate" would not have been counted as a quality problem. The high proportion of cases with Level I diagnostic errors yet judged to have had appropriate care suggested that these UPRO physician reviewers thought that diagnostic errors were not important if patient management was otherwise acceptable.

**Severity of diagnostic errors.** The Iliad review in the first experiment resulted in diagnostic errors that were judged to have caused more serious quality problems than those found by Iliad in the second experiment (see Table 20). Again, the largest contributing factor to this difference in severity was probably the gold standard review physicians used in the two experiments.

**The Iliad Review Process**

**Diagnostic Errors Not Detected by Iliad Review**

In Experiment I, the UPRO review did not detect any quality problems of a diagnostic nature. Although there were six cases found to have quality problems by both Iliad and UPRO review, the two review methods detected different types of problems (see Tables 8 and 9). However, in Experiment II, four diagnostic errors were found by the UCDS review but not by Iliad review. These four diagnostic errors were summarized in Table 15. The Iliad review of these four cases are examined here to account for the failure to flag these problems.

In the first case, the UCDS review physician judged that a diagnosis of pulmonary embolus might have been missed in a patient with deep vein thrombosis. Iliad did not pick up the problem because the patient had no respiratory problem on admission, but developed shortness of breath later on. These later findings were not entered into Iliad, as the experiment was planned with admission diagnostic workup. Complications developing later in the case were not used in the experiment.

In the second case, the UCDS physician believed that a diagnosis of acute gouty arthritis should have been made. Iliad did diagnose acute gouty arthritis, and this was
matched to the chart diagnosis of gouty arthropathy. It was not clear why the UCDS physician judged the case diagnosis inadequate.

In the third case, the UCDS physician reviewer decided that a diagnosis of azotemia was missed. The patient had mildly elevated blood urea and creatinine level on admission. Iliad did make the diagnosis of mild azotemia. However, this was considered an intermediate diagnosis in Iliad, and was used to contribute to the top-level diagnosis of acute renal failure, which did not reach a higher enough probability to trigger a flag. Since mild azotemia was not a top level diagnosis, Iliad did not display it in its diagnostic list, and thus no flag was triggered.

In the fourth case, the UCDS review physician judged that a diagnosis of pulmonary infection/fluid overload might have been missed in a patient with cough and, later on, shortness of breath. Iliad made the diagnosis of pneumonia with probability of 57%, not high enough to trigger a flag. This was because supporting evidence either developed later, like shortness of breath, or were not forthcoming, such as chest x-ray was not done, a problem remarked upon by the UCDS physician as well.

Therefore, it appeared that Iliad's failure to detect these four diagnostic errors was inevitable, given the current design of the experiment and the program.

**Missed Diagnoses Versus Unlikely Diagnoses**

It was interesting that a missed diagnosis appeared to be judged more seriously then an unlikely diagnosis in both Iliad reviews (see Tables 7 and 16). In addition, the four diagnostic errors found by the UCDS review were all problems of missed diagnoses (see Table 15). This might reflect the current prejudice, prevalent in the medical community, that doing more for a patient was usually better than doing less, a trend that might have contributed to the escalating health care costs. Part of this way of thinking might be due to medical-legal reasons, such as fear of malpractice suits, as well as cultural reasons, such as an attitude of "doing all one can" on the part of patients as well as physicians. The 1987 General Accounting Office study showed that among claims of diagnostic errors, 40% was
for an unlikely diagnosis, and 60% for a missed diagnosis. Despite literature showing problems due to unlikely diagnoses, it appeared likely that the present emphasis would still be on missed diagnoses.

Data Entry in Iliad Nurse Review: Interrater Reliability

The data entry of findings from a chart into Iliad, in the nurse review process, was explored for interrater reliability in Experiment II. Two UPRO nurses performed data entry for a sample of cases, and Cohen's kappa was used to measure interrater reliability between the two nurses, as well as between each nurse and the author, who performed all the data entry in this research. There was significant agreement between the two UPRO nurses, but not between each nurse and the author (see Figure 6). The lack of agreement between the nurses and the author was traced to a failure to enter important laboratory findings and procedure results on the part of the nurses. The reason for this failure might have been a lack of training on Iliad. In an effort to avoid introducing bias, each UPRO nurse was trained only an hour on Iliad, with no subsequent contact with the author during their data entry. The result might have been an unfamiliarity with the range of findings accepted by the program. It was possible, although not likely, that the nurses failed to realize the importance of laboratory and procedure findings, as those data not entered by the nurses were essential to diagnosis, such as electrocardiogram and cardiac enzymes in a patient with chest pain, or a chest x-ray in a patient with signs and symptoms suspicious of pneumonia, or thyroid enzymes for a diagnosis of hypothyroidism. Better training on Iliad might have resulted in a better agreement between the nurses and the author.

Diagnosis Matching in Iliad Nurse Review

After data entry into Iliad was completed, the Iliad diagnoses were compared to those made by the attending physician for each case. Cases were flagged when there was a discrepancy between the two diagnostic lists, according to the 80%/20% threshold criteria
explained earlier. Two methods of matching the diagnoses were explored in the two experiments.

The first experiment explored matching by medical terminology, using Cohen’s kappa to measure interrater reliability between this author and another physician not otherwise involved in this dissertation research. Kappa was statistically significant, indicating agreement beyond chance. Although an obvious problem of subjectivity remained, this matching method was intuitive and easy to perform. Because of the large variety of medical terminology, this matching by terminology might be made more objective by describing rules for nonexact matches. For instance, "ischemic cardiomyopathy" in Iliad would match to "coronary atherosclerosis" in the charts. These matching suggestions were not given to the physician asked to matched diagnoses in the interrater reliability exploration described earlier, to avoid bias.

The second experiment explored matching by ICD9 codes. Matching by using ICD9 codes solely was found to result in an unacceptably high number of false negative (an additional 70% of the original flags), due mainly to mistakes in coding in the charts. There was also the additional problem of missing or incomplete codes in the charts as well as in Iliad. It was thought that although matching solely by ICD9 codes was not cost-effective, ICD9 codes could be useful in helping to match some diagnoses.

Effect of Varying the Flagging Thresholds

The threshold criteria used to flag cases in the Iliad nurse review were set at 80%/20%, as explained earlier, based on empirical experience with Iliad. In Experiment II, the effect of tightening these threshold criteria was explored. Tightening referred to raising the threshold for flagging missed diagnoses from 80% upwards, and lowering the threshold for flagging unlikely diagnoses from 20% downwards. The results are shown in Table 19 and Figure 5. Generally, as the thresholds were tightened, there was a drop in Diagnostic Error Rate (DER). The effect on the nurse review True Positive Rate (TPR) was more variable. The thresholds of 80%/15% resulted in the highest DER and TPR,
followed by the 80%/20% thresholds used in the experiments, which had the same DER but a lower TPR. It thus appeared that the 80%/20% thresholds were well chosen for the experiments. The changes were generally not very large, reflecting a relative robustness when threshold criteria were used for flagging cases in Iliad nurse review.

**Gold Standard Physician Review: Interrater Reliability**

In Experiment I, each case flagged by the Iliad nurse review was referred to only one physician reviewer, to avoid overburdening the volunteer gold standard physicians. In Experiment II, due to financial constraints, each case flagged by Iliad nurse review was also referred to only one UPRO physician. Having only one Iliad gold standard review per case was recognized as a weakness in the experiment design, since it was recognized that medical opinions vary greatly, as evidenced by the large variation in medical practice documented in the medical literature.\(^8,9,169,170\) Although the PRO review had not been specifically evaluated as to the interrater reliability among review physicians, PROs were aware that review results varied from PRO to PRO. Physicians had also expressed dissatisfaction with the poor interrater agreement in PRO peer review.\(^6,25,171,172\)

Although interrater reliability within a group of gold standard physician reviewers, namely, University of Utah physicians in Experiment I and UPRO physicians in Experiment II, was not studied, a sample of cases in Experiment II were reviewed by gold standard physicians from both of these two groups. The same four University of Utah physicians, who performed the gold standard Iliad review in Experiment I, were asked to independently review 15 cases flagged by Iliad nurse review in Experiment II. Their review results were compared to that achieved by the UPRO physicians. Interrater reliability between the two groups, as measured by Cohen's kappa, was statistically not significant. Hence, any agreement in Iliad gold standard review between University of Utah physicians and UPRO physicians was only by chance.\(^{164,165}\) This lack of agreement between the two groups of gold standard reviewers was not unexpected, given the previously discussed experimental results and the differences between the two groups. It
was discussed earlier that the type or characteristics of gold standard physicians would affect the results of Iliad review, and the insignificant kappa between the gold standard reviews performed by University of Utah physicians and UPRO physicians supported this suggestion.

Next, a sample of cases in Experiment II went through Iliad review by a unique physician who was a specialist (pulmonary physician, but not the same one in Experiment I), had teaching responsibilities, was involved in Iliad research, and also was on UPRO's panel of review physicians. Thus this unique physician had background characteristics common to both groups of gold standard physicians. Interrater reliability, as measured by Cohen's kappa, was calculated between this unique physician and the group of University of Utah physicians, as well as between this unique physician and the group of UPRO physicians. The results were represented in Figure 7. Both kappas were positive and statistically significant, indicating more than chance agreement between this unique physician and University of Utah physicians, as well as between this unique physician and UPRO physicians. Yet the Cohen's kappa between University of Utah and UPRO physicians was earlier calculated to be statistically not significant. Thus the unique physician, sharing background with both groups of physicians, appeared to achieve agreement with each of the two groups, whereas the two groups, with no common background, were not able to agree with each other. This evidence indirectly supported the conclusion that the type of gold standard reviewers strongly affected the performance of Iliad in diagnostic error detection.

Cost-Benefit Estimation of Iliad Review

A cost-benefit estimation was made for using Iliad review in both experiments (see Tables 10 and 17). Iliad review was not intended to replace the present manual PRO review, nor the future UCDS review. Since Iliad review was designed to detect diagnostic errors, and the PRO and UCDS reviews concentrated on treatment, documentation and discharge problems, it was logical to propose combining Iliad review with either of these
review methods. Therefore, the cost-benefit estimation in either experiment was for the combined review. It could be seen that adding Iliad review to either the manual PRO review or the UCDS review resulted in more quality problems being detected, at a lower cost per problem, although the net cost increased predictably.

The cost of Iliad review can be split between the nurse review and the gold standard physician review. The nurse review cost was largely that of data entry. Some amount of time was required in the two experiments to tabulate (in a spreadsheet) and match diagnoses, because this step was delayed until all data entry in an experiment was completed, then performed for all the cases together. In a real review, the Iliad diagnoses and attending physician diagnoses could be compared for each case directly after data entry, which would take very little time, especially when compared to the data entry. The time needed for data entry would, in turn, depend on the version of the Iliad program and the capabilities of the computer used. The faster program and computer used in the second experiment resulted in less time per case.

If physician reviewers were to be reimbursed on a per-case basis (as by UPRO), the time needed for the Iliad gold standard review would not be important. The cost of physician review for Iliad would then depend on the True Positive Rate of the Iliad nurse review, which would decide the number of unnecessary physician reviews due to false positives. However, if physician reimbursement was to be time-dependent, the Iliad gold standard review would cost relatively little, as the physicians required much less time than in the PRO or UCDS review.

Because the UPRO review in Experiment I was not very expensive ($107 per problem found), adding Iliad review lowered the cost per problem only 10%. However, the number of problems detected increased by 40%, to 39 out of 100 cases reviewed. The UCDS review in the second experiment was noticed to be very much more expensive than the manual UPRO review ($1,773 per problem found). Adding Iliad review just to the cases already flagged by the UCDS nurse review dropped the cost per problem to $684,
only 40% of that from UCDS review alone, while tripling the number of problems detected to 31%. Adding Iliad review to either the PRO manual or UCDS review thus appeared to be cost-effective, despite the increase in net cost. For the manual UPRO review, adding Iliad review would increase the net cost by 24%. In the case of adding Iliad to the UCDS review, the cost of UCDS review was so high that the added net cost would be less than 9%.

The severity of the quality problems detected with the combined review was higher in the first experiment, but lower in the second. This was expected since Iliad review detected more serious errors in Experiment I, but less serious errors in Experiment II. The reason for this difference in severity of problems detected was felt to be the different types of gold standard reviewers in the two experiments.

Experiment II also explored the possibility that diagnostic errors might be more common in cases suspected of having management problems by the UCDS nurse review. It was believed that it might be more cost-effective then to add Iliad review only to this group of cases. However, the results of the experiment did not support this hypothesis. There was no significant difference found between the number of diagnostic errors in the group of cases flagged by the UCDS nurse review and the number of diagnostic errors in the group thought problem-free. Adding Iliad review to all the cases, compared to adding Iliad review to only those cases already flagged by the UCDS nurse review, would increase the net cost by less than 8%. The cost per quality problem would drop 16% to $572, and the number of problems detected would increase to 40%. Thus, adding Iliad review to all cases reviewed by UCDS would also be cost-effective.
CONCLUSION

Research Hypothesis

Iliad review was shown to detect more quality problems caused by diagnostic errors, compared to the present PRO review employed by UPRO for Medicare hospitalizations, as well as the UCDS review planned by HCFA for future implementation. Iliad review was also shown to have a higher true positive rate than either of these two Medicare inpatient review methods.

Recommendations

It has been shown in this dissertation that Iliad review achieved its objective of detecting diagnostic errors not found by current and future PRO review methods for the Medicare inpatient population. Iliad review thus complemented and supplemented the PRO reviews. Combined with the manual PRO review or the computerized UCDS review, Iliad review was shown to be cost-effective.

To used Iliad review on a widespread basis, as part of quality review activities, would require overcoming certain obstacles. The issues to be addressed could be divided into two groups. First would be the performance or technical issues. Certain improvements and additions could be made to the Iliad program to result in a version specifically designed for quality review. Ideally, the nurse reviewer should need to perform only two tasks: data entry (until technology has advanced to the stage of automated retrieval from a computer-based patient record), and confirming or overruling the flagging recommendation made by Iliad. Thus, the review version of Iliad would ask for the chart diagnoses, to be compared to the diagnoses it made for the case, using pre-determined threshold criteria and matching rules similar to those explored in this
dissertation, perhaps with ICD9 code matching as backup. Diagnoses flagged as missing or unlikely by Iliad could be further reviewed for confirmation by the nurse, using information from the chart, before referring the case to physician review. Such a program would reduce nurse review time and false positive flags, thus improving cost-effectiveness. Also, it would allow for flexibility as well as objectivity.

Apart from the above, an important obstacle to be overcome, before widespread use of Iliad review could be accepted, would be the attitude of health care providers. This dissertation has shown that the type or characteristics of gold standard physician reviewer affected the performance of Iliad. It is believed by the author that for any quality review activity to be effective, whatever the method used, the people involved (both the reviewer and the care providers being reviewed) must believe in it and want it. The PRO program was not thought to be a success because

> despite two decades of inspired leadership in the medical community in the quality assurance and peer review movement, many physician remain suspicious of and hostile to PRO activities, continuing to perceive it at one and the same time as intrusive, arbitrary, and punitive - and fundamentally irrelevant to improving quality of care.⁶

When there is strong resistance, or when the quality assurance or review activity is regarded as a regulatory nuisance, instead of learning from feedback or benefiting in other ways, the care provider being reviewed is likely to dismiss the quality of care information provided. When the quality review method proposed involved an expert system like Iliad, there is the additional barrier formed by the misunderstanding many health care providers have of the role of computers in health care. Clinicians and quality assurance practitioners might feel that computers have no place judging or dictating the clinical practice of a physician (personal experience at a recent national quality assurance conference). Iliad review is not designed to replace human judgment, but to act as a tool, just like a textbook can be a tool. However, the additional persuasion needed to convince care providers to accept the role of Iliad, together with that needed to overcome the hostility already present against quality review, will be a formidable undertaking.
The use of Iliad review thus needs to be considered in the larger context of quality review as a whole. Quality review is a highly controversial topic. However, it is unlikely to be completely dismissed, as evidenced by discussions within the health care reform. This dissertation argues that, if quality review must be performed, an expert system tool such as Iliad can be logical and cost-effective. Also, this dissertation evaluated Iliad review by comparing it to the current and future PRO reviews for Medicare inpatients. It is likely that Iliad can be used in other quality review activities, for instance, in an internal quality assurance program in a hospital. Iliad has other features, not described in this dissertation, that enabled it to be useful as a medical education tool. For example, one of Iliad's features is a case simulation mode, in which Iliad acts like a fictitious patient, and the user is expected to reach a diagnosis on Iliad's problem. Iliad will start with presenting a chief complaint, and the user is expected to ask relevant questions on the medical history, followed by physical examination and tests. For instance, the user might ask for an abdominal palpation, and Iliad might reply "tenderness in the right lower quadrant." The user is graded on the diagnosis reached eventually, as well as the cost-effectiveness of the path taken to the diagnosis. This case simulation function of Iliad has been used effectively in teaching medical students diagnostic skills, and can be combined with its quality review function to form a complete quality assurance package with both review and remedial education.

In conclusion, a diagnostic expert system such as Iliad can be used effectively as a screening tool to help PRO review nurses flag diagnostic errors. It is clear from the data presented that diagnostic error detection in individual chart review by individual physician reviewers is not reproducible. On the other hand, expert systems such as Iliad can be used to create and tune a diagnostic model, based on the consensus opinion of experts from a variety of backgrounds. This diagnostic model can then act as a standard, with which the diagnosis made by the attending physician can be compared. Chart review by individual physicians can then be substituted by periodic sampling to evaluate and fine-tune the
diagnostic model, again by expert consensus opinion. This would drastically reduce the cost and inconsistency of quality review.
REFERENCES


