A DECISION SUPPORT SYSTEM FOR THE DIAGNOSIS AND MANAGEMENT OF PNEUMONIA PATIENTS

by

Dominik Aronsky

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

May 2001

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a dissertation submitted by

Dominik Aronsky

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

[redacted] Chair: Peter J. Haug

[redacted] Reed M. Gardner

[redacted] R. Scott Evans

[redacted] [redacted]

[redacted] [redacted]

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

FINAL READING APPROVAL

To the Graduate Council of the University of Utah:

I have read the dissertation of Dominik [redacted] 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 committe and is ready for submission to The Graduate School.

12-15-00

Date

[redacted]

Approved for the Major Department

[redacted]

Reed M. Gardner
Chair

Approved for the Graduate Council

[redacted]

David S. Chapman
Dean of The Graduate School

ABSTRACT

Diagnostic decision support systems have a long tradition in medicine, but have not typically been integrated with clinical information systems. I describe five studies that were performed during the development, implementation, and evaluation efforts for a real time diagnostic decision support system. The system's objective was to automatically identify patients likely to have pneumonia. The real time system used only data routinely collected in the computerized patient record during a patient's encounter in the emergency department. The automatic identification of pneumonia patients was used to initiate the computerized calculation of a pneumonia risk assessment instrument.

The first study describes the development of the diagnostic system using an emergency department data set from the HELP clinical information system at LDS Hospital, Salt Lake City, Utah. The second study describes the implementation of the system and illustrates the operational functions. The third study examined whether data from the HELP System could be used for the computerized evaluation of the pneumonia risk assessment instrument. This study showed that computerized evaluation generated an accurate risk class for 86 percent of hospitalized pneumonia patients. The fourth study reports the design and planning of the system's prospective evaluation in a clinical environment. The paper addresses important issues that influenced the study design, such as verification bias and disease prevalence. Different study designs were discussed with respect to the targeted users and the clinical setting, and a feasible approach for the creation of a valid gold standard diagnosis for pneumonia was proposed. The final study describes the system's prospective clinical evaluation. During a 5-month study period, the system computed a probability of pneumonia in real time for 10,828 patients, of whom 265 patients had pneumonia. The diagnostic accuracy, determined by the area under the receiver operating characteristic curve, was 0.942.

In summary, this project describes the development, implementation, and evaluation of a fully automatic, real time, diagnostic system that is integrated with a clinical information system. The system can be used to initiate pneumonia specific risk assessment tools and guidelines and to support the computerized guideline implementation for delivering recommendations at the point of patient care.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge my principal advisor Peter J. Haug, MD, who guided me through my learning and research experience as I started and continued my medical informatics path. Special thanks go to Reed M. Gardner, PhD, who was on my committee and supported me with enthusiasm and criticism during, and I hope beyond, my studies. I am thankful for the advice and the teaching of my other committee members: Scott R. Evans, PhD, Clifford J. Drew, PhD, and Michael J. Lincoln, MD.

I am thankful to the many clinicians with whom I had the opportunity to collaborate on different projects: Nathan C. Dean, MD, took me on his journey exploring new ideas and approaches for community-acquired pneumonia; Mary Suchyta, MD, supported me with endless time for reviewing charts and enabling many of my projects; the emergency department nurses and physicians supported me during my project.

Thanks to the medical informatics staff at LDS Hospital who had the time and patience to support me with advice at any time and without whom my project would not have been as much fun. I am thankful for the friendship of Wendy Webber Chapman, PhD, and Marcelo Fiszman, MD, with whom I had interesting discussions and joint projects.

I would like to thank my wife Eva and my children Aline and David for all their smiles, laughter, and patience during my studies.

I was financially supported by an MD-PhD grant from the Swiss National Science Foundation (31-44406.95). The final evaluation of the decision support system was supported by a grant from the LDS Hospital Deseret Foundation (# 311).

Permission to reprint the two papers published in the Proceedings of the American Medical Informatics Association and the paper published in the Journal of the American Medical Informatics Association was granted by the medical publishers Hanley & Belfus, Inc., Philadelphia (PA) on September 25, 2000 and September 29, 2000.

CHAPTER 1

INTRODUCTION

1.1 Clinical Information Systems

Clinical information systems are installed in hospitals with increasing frequency and are becoming the standard medium for managing patient information. One of the remaining goals for clinical information systems is the creation of a single, lifelong patient record that can be accessed at any time, independent of the patient's or health care provider's location. Because clinical information systems are expensive to develop, install, and, in particular, to maintain, a financial return on investment is expected. The basic functions of a clinical information system are collecting, storing, and reporting patient information. A clinical information system that uses only the basic functions does not differ from the paper-based patient chart with the exception of time and location independent access to patient information. Added value and improved patient care can be expected from clinical information systems that are able to utilize the large amount of stored information. Health care has only started to mine and exploit the rich data sources that clinical information systems offer. "Putting the information to work" in the form of decision support systems and using the data stored in clinical information systems has already demonstrated an impact on patient care (1-4).

1.2 Decision Support Systems

Decision support systems build on top of the three basic functions of a clinical information system. Integrating decision support with clinical information systems allows such systems to interact with health care providers in a transparent way during the collection and reporting of information. Decision support systems can be categorized in several different ways. A functional view differentiates between systems that alert, critique, suggest, or are used for quality assurance (5).

Alerting systems use patient data to decide whether an alert needs to be generated. Alerting systems can be driven by data or by time. Data-driven systems process information as clinical data become available and time-driven systems process information in specified time intervals. An example for a data-driven alerting system is a laboratory system that is activated whenever new laboratory results become available (6). An example for a time-driven system is an immunization reminder system that is activated at specific times and generates reminders for the adult immunizations (7).

Critiquing systems are triggered by user input. They combine the user input with clinical data and provide information about criteria that were established from institutional policies, clinical guidelines, or federal agencies. An example for a critiquing system is a blood-order monitor that provides a user with feedback if specified criteria for ordering a blood product are not met (8).

Systems that suggest an action are typically initiated by the user and provide a recommendation based on clinical data from an information system and might ask the user for additional data. An antibiotic assistant is an example for a suggesting system where a user requests suggestions for the selection of an empiric antibiotic regimen (4).

Quality assurance systems are used to examine and provide information about clinical processes. An example for a quality assurance system could be the reporting of the time interval between the end of surgery and the extubation for patients with coronary-artery-bypass-graft surgery.

A different view for categorizing decision support systems classifies systems based on the applied inference mechanism (9). Algorithms for decision support systems include different representations such as pathophysiologic models, statistical models, pattern recognition, case-based reasoning, or symbolic reasoning. Symbolic reasoning or knowledge-based systems typically consist of an inference mechanism and a knowledge base. Frequently applied inference mechanisms for knowledge-based systems are rules, heuristics, frames, decision trees, and probabilities. Nonsymbolic reasoning systems or systems that do not rely on a knowledge base are artificial neural networks or genetic algorithms that can be applied in combination with artificial neural networks.

1.3 Diagnostic Systems

Diagnostic decision support systems have long been a favorite area of research for clinicians and medical informaticists. Investigators have recognized a clinician's limitation of evaluating a large amount of data for making medical diagnoses. In an attempt to integrate large amounts of data the human brain is subject to bias that influences the medical decision making process (10, 11). With the availability of powerful computers, researchers hoped that the exact, mathematical support, which computers can provide, might help clinicians in the diagnostic reasoning process.

Diagnostic decision support systems have attracted researchers since the availability of computers. One of the first applications was created by Homer R. Warner, MD, PhD, professor emeritus and a pioneer in medical informatics at LDS Hospital/University of Utah, who investigated the application of Bayes' theorem to the diagnosis of heart diseases in 1961 (12). During the last four decades a myriad of diagnostic systems have been developed (13). During the last decade fast and powerful personal computers gave researchers access to computational resources that allow the examination of large databases and the performing of complex data mining tasks. Consequently, the number of diagnostic decision support systems has increased even more.

Large-scale, general-purpose diagnostic systems cover many medical diseases. Examples of large-scale systems include QMR, DxPlain, Iliad, and Meditel (14-17). The work involved in developing general-purpose diagnostic systems is huge and lasts multiple years. The performance of the four above-mentioned, large-scale systems was evaluated in a comparative study and showed that they had comparable diagnostic accuracy even though each of the systems used a different method for making inferences (18).

More focused diagnostic systems, targeting a smaller set of diseases within a specialty discipline, were developed more frequently than large-scale systems. Mentioned here are the few focused systems that have undergone a clinical evaluation. In Leeds, England, a simple Bayesian system, a precursor of Bayesian networks, was developed for the diagnosis of acute abdominal pain (19). The system was evaluated in several studies and different hospitals (20-22). A logistic regression model for the identification of acute cardiac ischemia included seven clinical variables and the help of a hand held calculator to obtain a patient's probability of having an acute cardiac ischemia (23). The logistic regression model was evaluated in a prospective, multicenter study and demonstrated a significant reduction in coronary care unit admissions (24). A similar model predicting the probability of acute cardiac ischemia was integrated into an electrocardiography machine and, in a multicenter, controlled study, demonstrated a reduced hospitalization rate for patients without acute cardiac ischemia (25). An artificial neural network for the diagnosis of acute myocardial infarction was prospectively evaluated and outperformed the participating physicians (26, 27). Growing cell structure networks are a new method that provide a graphical rather than a numerical output (28). A growing cell structure network was developed and tested for the cytodiagnostic identification of breast carcinoma. Pathfinder is a Bayesian network developed for the diagnosis of lymph node diseases (29, 30). The diagnostic performance of Pathfinder was at least as accurate as the pathologist who helped develop the system (31). The system was subsequently evaluated in an independent study and was judged to be a valuable tool supporting pathologists in the diagnosis of lymph node pathologies (32).

Numerous other diagnostic systems have been developed using a variety of artificial intelligence algorithms including logistic regression, artificial neural networks with and without enhancement by genetic algorithms, Bayesian networks, and decision trees. However, the majority of systems were developed in a laboratory setting generally using only a single data set for the development and the evaluation of the system. Unfortunately, very few systems have been evaluated in a prospective clinical evaluation.

One of the major drawbacks of previously reported diagnostic decision support systems was their lack of integration into a clinical information system. The stand-alone nature of diagnostic systems jeopardizes their routine application in a busy clinical setting. Clinicians have to interact directly with stand-alone systems and are required to enter a variety of patient information. Entering data in a stand-alone system has the disadvantage that some of the patient information might already be available in the clinical information system resulting in redundant data entry. The additional time spent interacting with the computer is considerable and represents time not spent with the patient.

1.4 Behavioral Bottleneck

Many clinical guidelines and predictive instruments are complex and time consuming to follow. Such complexity is often required to give disease- and patient-specific recommendations. Even though the guidelines and predictive instruments have demonstrated a positive effect on patient care, only the simplest ones, such as the Ottawa ankle rule (33), are used for routine patient care. Despite the ability to computerize guidelines and predictive instruments clinicians remain the "gatekeepers" for applying them. Clinicians are responsible for identifying patients with a particular disease, going through the process of physically locating the paper-based or computerized guideline, and following the all required steps to obtain a patient-specific recommendation.

There are several reasons, such as time constraints, workflow integration, logistical processes, and psychological reasons, why clinicians often keep the "gate shut" when it comes to applying guidelines. I termed the sum of reasons why clinicians keep the guideline gate shut a "behavioral bottleneck." The behavioral bottleneck describes a situation in which patient information and guidelines or predictive instruments that have been demonstrated to be effective are available to clinicians but are not applied.

An example of the behavioral bottleneck is the clinical application of the Pneumonia Severity of Illness Index (34). The Pneumonia Severity of Illness Index is one of the best evaluated risk assessment tools currently available and meets almost all recommendations for the development and evaluation of risk prediction tools (35, 36).

The severity index was developed as a logistic regression (37). Because logistic regression models are not practical for routine clinical use, the investigators simplified and transformed the logistic regression model into a simpler scoring algorithm that would be easier to apply in a clinical setting. During the transformation of the statistical to the clinically applicable model, several simplifications were necessary. The transformation of a statistical model into a simple algorithm often comes at the cost of losing information and sacrificing predictive accuracy for simplicity. The transformed scoring algorithm requires twenty variables routinely collected during the patient's encounter. However, it appears that the scoring algorithm retains a level of complexity that keeps clinicians from applying the clinical version of the severity index for managing pneumonia patients.

Clinical information systems have the computational power, speed, and data required to calculate the Pneumonia Severity of Illness Index during the patient's encounter. The computerization of the severity index is possible both in its complex logistic regression version and its simpler, clinically applicable scoring version. The computerization of the severity index, however, is not sufficient to guarantee its clinical application because clinicians must also initiate the process by identifying patients for whom they wish to apply the risk assessment instrument. Thus, the behavioral bottleneck remains present independent of whether the instrument is available in a paper- or computer-based format. If a diagnostic system can identify pneumonia patients automatically, the computerization of the severity index could be initiated without user interaction and would make the results of the severity index available to physicians in an unsolicited way.

1.5 Objective

The objective for this project was to overcome the behavioral bottleneck. For this purpose I developed, implemented, and evaluated a real time diagnostic system that has the ability to automatically identify patients likely to have pneumonia. To provide clinicians with guideline recommendations at the point of care, the diagnostic system had to be accurate and ideally would not require health care providers to enter additional data.

1.6 A Diagnostic System for Identifying Patients with Pneumonia

In this dissertation I report the different phases of developing, implementing, and evaluating a decision support system for the real time identification of patients likely to have pneumonia in an emergency department setting. The decision support system was integrated with the clinical information system and uses only routinely available data. It does not require health care providers to enter additional data but operates with patient information collected as part of the regular computerized charting and documentation. The system has a diagnostic and a management component. The diagnostic component is responsible for identifying patients likely to have pneumonia. The management component consists of the Pneumonia Severity of Illness Index, which is a risk assessment instrument for pneumonia patients. The severity index yields a mortality risk class of 1 to 5, where 1 indicates a low risk and 5 a high risk of mortality. The developers of the severity index suggested that the risk assessment instrument is applicable for the identification of low risk patients who could be treated on an ambulatory basis.

1.6.1 Development

Chapter 2 describes the development phase of the system's diagnostic component. I developed a Bayesian network by applying a historical data set of more than 32,000 patients from the emergency department of LDS Hospital. Bayesian networks (38, 39) are probabilistic representations and follow the mathematics of Bayes' theorem (40). Bayesian networks consist of nodes representing random variables. A table is attached to each node containing the conditional probabilities used for updating the joint probability. The nodes in the Bayesian network are connected through directed links. The links model the conditional dependency among the network nodes and represent the network structure. During the development phase more than 100 different network structures were trained and evaluated in an attempt to obtain an accurate and parsimonious model.

1.6.2 Implementation

Chapter 3 explains the implementation and the integration phase of the decision support system. The decision support system was integrated with the HELP System (41, 42), which is the clinical information system at LDS Hospital, and the radiology information system. The system was implemented as a client-server application that used the HELP System as the server and the decision support system as the client. Information from the diagnostic component and the management component was made available to the emergency department staff on the clinical information system as well as on a dedicated computer in the emergency department. The HELP System displayed only the probability of pneumonia, calculated from the diagnostic component and the patient's

Pneumonia Severity of Illness risk class. The dedicated computer gave the emergency department staff access to more detailed information about a specific patient.

The diagnostic component allowed interested users to examine the reasoning process in depth and to obtain an explanation of how the current pneumonia probability was calculated. For that purpose the temporal course of the pneumonia probability, the variables included in the probability calculation, and the resulting probability changes were displayed. The management component gave the emergency department staff access to the detailed calculations of the Pneumonia Severity of Illness Index. All the variables involved in the risk assessment of pneumonia patients and the respective risk scores were also available for detailed examination.

1.6.3 Data Quality and Completeness

Chapter 4 reports on a study that examined whether the data in the clinical information system was accurate enough to be used in the management component of the decision support system. The management component applied 20 clinical variables of which 19 were routinely collected during routine patient care and computed a risk class (1 to 5) indicating a patient's estimated risk of mortality. I performed a retrospective chart review of 241 hospitalized patients to test whether data originating from the clinical information system could be used for the automatic, real time calculation of the Pneumonia Severity of Illness Index. The study's objective was to compare the quality of available, computerized data alone with a reference standard consisting of all data available in the paper chart and the clinical information system at the time of the emergency department encounter. Real time, computerized calculation of the Pneumonia Severity of Illness Index was accurate for 86 percent of hospitalized patients.

To assess the data quality of a clinical information system for the real time calculation of a predictive instrument or a clinical guideline is important. Without knowing the level of data quality even a well-designed decision support system might fail in a real world implementation because physicians might not trust the displayed information.

1.6.4 Study Design

Chapter 5 consists of a design paper reporting the planning phase for the prospective clinical evaluation of the decision support system. Prospective clinical evaluation studies of decision support systems remain infrequent. In this paper important issues were addressed that might influence the study design, such as verification bias, the influence of disease prevalence, and the creation of a valid gold standard for the diagnosis of pneumonia. The advantages and disadvantages of different study designs are explored with respect to the targeted users and the clinical setting.

1.6.5 Clinical Evaluation

Chapter 6 reports the results of the prospective clinical evaluation of the diagnostic component in the emergency department at LDS Hospital. During a 5-month period (November 12, 1999 to April 15, 2000) the system computed a probability of pneumonia for 10,828 patients. A gold standard diagnosis was established reviewing all study patients. The review applied a three-step process. The initial review step included the application of a set of criteria to all study patients. Study patients not meeting any of the criteria were considered not having pneumonia. The second step included a review of the emergency department report and the reports of the radiology exams by five internal medicine physicians. Patients without evidence of pneumonia were excluded from further review. In the last review step two pulmonary and critical care physicians reviewed a patient's chart and radiology images and established a diagnosis. A third pulmonary and critical care physician reviewed patients whose pneumonia diagnosis was not established by both initial reviewers. The majority vote of the three physicians decided on the final pneumonia diagnosis. The review process found a gold standard diagnosis of pneumonia for 265 patients among the 10,828 study patients.

1.7 References

1. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA. 1998;280:1339-46.

2. Johnston ME, Langton KB, Haynes RB, Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. A critical appraisal of research. Ann Intern Med. 1994;120:135-42.

3. Balas EA, Austin SM, Mitchell JA, Ewigman BG, Bopp KD, Brown GD. The clinical value of computerized information services. A review of 98 randomized clinical trials. Arch Fam Med. 1996;5:271-8.

4. Evans RS, Pestotnik SL, Classen DC, Clemmer TP, Weaver LK, Orme JF Jr, et al. A computer-assisted management program for antibiotics and other antiinfective agents. N Engl J Med. 1998;338:232-8.

5. Haug PJ, Gardner RM, Evans RS. Clinical Decision Support Systems. Berner ES (ed.) Springer Verlag NewYork, Inc. 1999;77-103.

6. Bradshaw KE, Gardner RM, Pryor TA. Development of a computerized laboratory alerting system. Comput Biomed Res. 1989;22:575-87.

7. McDonald CJ, Hui SL, Tierney WM. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. MD Comput. 1992;9:304-12.

8. Gardner RM, Golubjatnikov OK, Laub RM, Jacobson JT, Evans RS. Computer-critiqued blood ordering using the HELP system. Comput Biomed Res. 1990;23:514-28.

9. van Bemmel JH, Musen MA, Handbook of Medical Informatics. Springer Verlag Heidelberg, 1997.

10. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science. 1981;211:453-8.

11. Eddy DM. Practice policies: Where do they come from? JAMA. 1990;263:1265-75.

12. Warner HR, Toronto AF, Veasey LG, Stephenson RA. Mathematical approach to medical diagnosis. JAMA. 1961;177:75-81.

13. Miller RA. Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc. 1994;1:8-27.

14. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. JAMA. 1987;258:67-74.

15. Miller RA, Pople HE Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982;307:468-76.

16. Warner HR Jr. Iliad: moving medical decision-making into new frontiers. Methods Inf Med. 1989;28:370-2.

17. Waxman HS, Worley WE. Computer-assisted adult medical diagnosis: subject review and evaluation of a new microcomputer-based system. Medicine (Baltimore). 1990;69:125-36.

18. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994;330:1792-6.

19. Horrocks JC, McCann AP, Staniland JR, Leaper DJ, De Dombal FT. Computer-aided diagnosis: description of an adaptable system, and operational experience with 2,034 cases. Br Med J. 1972;2:5-9.

20. De Dombal FT, Leaper DJ, Horrocks JC, Staniland JR, McCann AP. Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians. Br Med J. 1974;1:376-80.

21. Staniland JR, Clamp SE, de Dombal FT, Solheim K, Hansen S, Ronsen K, et al. Presentation and diagnosis of patients with acute abdominal pain: comparisons between Leeds, U.K. and Akershus county, Norway. Ann Chir Gynaecol. 1980;69:245-50.

22. McAdam WA, Brock BM, Armitage T, Davenport P, Chan M, de Dombal FT. Twelve years' experience of computer-aided diagnosis in a district general hospital. Ann R Coll Surg Engl. 1990;72:140-6.

23. Pozen MW, D'Agostino RB, Mitchell JB, Rosenfeld DM, Guglielmino JT, Schwartz ML, et al. The usefulness of a predictive instrument to reduce inappropriate admissions to the coronary care unit. Ann Intern Med. 1980;92:238-42.

24. Pozen MW, D'Agostino RB, Selker HP, Sytkowski PA, Hood WB Jr. A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. A prospective multicenter clinical trial. N Engl J Med. 1984;310:1273-8.

25. Selker HP, Beshansky JR, Griffith JL, Aufderheide TP, Ballin DS, Bernard SA, et al. Use of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms suggestive of acute cardiac ischemia. A multicenter, controlled clinical trial. Ann Intern Med. 1998;129:845-55.

26. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med. 1991;115:843-8.

27. Baxt WG, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. Lancet. 1996;347:12-5.

28. Walker AJ, Cross SS, Harrison RF. Visualisation of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique. Lancet. 1999;354:1518-21.

29. Heckerman DE, Horvitz EJ, Nathwani BN. Toward normative expert systems: Part I. The Pathfinder project. Methods Inf Med. 1992;31:90-105.

30. Heckerman DE, Nathwani BN. Toward normative expert systems: Part II. Probability-based representations for efficient knowledge acquisition and inference. Methods Inf Med. 1992;31:106-16.

31. Heckerman DE, Nathwani BN. An evaluation of the diagnostic accuracy of Pathfinder. Comput Biomed Res. 1992;25:56-74.

32. Nathwani BN, Clarke K, Lincoln T, Berard C, Taylor C, Ng KC, et al. Evaluation of an expert system on lymph node pathology. Hum Pathol. 1997;28:1097-110.

33. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Reardon M, et al. Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. JAMA. 1993;269:1127-32.

34. Fine MJ, Auble TE, Yealy DM, Hanusa BH, Weissfeld LA, Singer DE, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med. 1997;336:243-50.

35. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985;313:793-9.

36. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA. 1997;277:488-94.

37. Fine MJ, Hanusa BH, Lave JR, Singer DE, Stone RA, Weissfeld LA, et al. Comparison of a disease-specific and a generic severity of illness measure for patients with community-acquired pneumonia. J Gen Intern Med. 1995;10:359-68.

38. Jensen FV. An introduction to Bayesian Networks, Springer-Verlag New York Inc., New York, 1997.

39. Pearl J. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, California, 1988.

40. Szolovits P. Uncertainty and decisions in medical informatics. Methods Inf Med. 1995;34:111-21.

41. Kuperman GJ, Gardner RM, Pryor TA: HELP: A Dynamic Hospital Information System. New York: Springer-Verlag, 1991.

42. Gardner RM, Pryor TA, Warner HR. The HELP hospital information system: update 1998. Int J Med Inf. 1999;54:169-82.

CHAPTER 2

DIAGNOSING COMMUNITY-ACQUIRED PNEUMONIA WITH A BAYESIAN NETWORK

Dominik Aronsky, MD, Peter J Haug, MD

Proc AMIA Symp. 1998;:632-6.

(reprinted with permission from Hanley & Belfus, Inc.)

Diagnosing Community-Acquired Pneumonia with a Bayesian Network

Dominik Aronsky, MD and Peter J. Haug, MD
Dept. of Medical Informatics, LDS Hospital/University of Utah, Salt Lake City, Utah

We present the development and the evaluation of a Bayesian network for the diagnosis of community-acquired pneumonia. The Bayesian network is intended to be part of a larger decision support system which assists emergency room physicians in the management of pneumonia patients. Minimal data entry from the nurse or the physician, timely availability of clinical parameters, and high accuracy were requirements we tried to meet. Data from more than 32,000 emergency room patients over a period of 2 years (June 1995–June 1997) were extracted from the clinical information system to train and test the Bayesian network. The network performed well in discriminating patients with pneumonia from patients with other diseases. The Bayesian network achieved a sensitivity of 95%, a specificity of 96.5%, an area under the receiver operating characteristic of 0.98, and a predictive value positive of 26.8%. Our feasibility study demonstrates that the proposed Bayesian network is an appropriate method to detect pneumonia patients with high accuracy. The study suggests that the proposed Bayesian network may represent a successful component within a larger decision support system for the management of community-acquired pneumonia.

Introduction

Community-acquired pneumonia (CAP) is the sixth leading cause of mortality in the US and the leading cause of death in patients with infectious diseases.1 The cost of CAP is estimated to be $4 billion per year.2 The diagnosis and the management of CAP involves much uncertainty when a patient presents to the Emergency Room (ER). At this point, however, important decisions about the empiric antibiotic selection and the admission to the hospital have to be made. Making decisions under uncertainty results in practice variation.

To reduce practice variation guidelines for the management of patients with CAP have been developed.[3,4] One of them has been successfully implemented in the medical delivery systems of Intermountain Health Care.[4] The guideline is paper based and requires additional time to be filled out. Therefore the physicians' compliance varies. Computerizing the guideline may increase the compliance. However, a computerized guideline may only be successful if a sensitive and specific trigger mechanism that accurately identifies patients with CAP is present.

As the quality of computerized patient records improve, decision support systems represent a promising method to improve patient outcomes and cost-effectiveness. Most of the current decision support systems are rule-based. Probabilistic methods such as Bayesian networks still need to demonstrate their value and applicability in an integrated clinical environment.

A Bayesian network (BN) is a graphical representation that is based on probability theory, primarily on Bayes' theorem.[5,6] A BN is a directed acyclic graph with nodes, arcs and tables. Each node represents an uncertain variable and is associated with a table representing a probability distribution. The arcs describe the probabilistic dependencies among the variables. A joint probability for the variables can be calculated by propagating and updating the probabilities through the BN. An efficient algorithm to propagate probabilities and calculate a joint probability has been developed[7], but the computational time grows exponentially as the number of nodes in the BN increases.[8] Although BNs are becoming more manageable with the increasing power of computers, major problems remain. The estimation of the conditional probabilities by literature review or with the help of domain experts is tedious and time consuming. In particular, the probabilities of findings in the population without the target disease are difficult to assess. Although clinical databases can potentially provide accurate probabilities, they have not been deployed for the development of a BN, because they often lack the required detail.

CAP is a good candidate for the probabilistic nature of a BN because uncertainty is involved in both the diagnosis and in the management of the disease. The physician encounters much variation in symptoms, findings, and laboratory and blood gas test results. Even in the chest x-ray, which is considered the gold standard for the diagnosis of CAP, the interpretations may vary. The sputum and blood cultures take one or two days to be completed and are often negative. At the time the reports become available, most important decisions have already been made.

In our feasibility study we present the development and the evaluation of a BN for the diagnosis of community-acquired pneumonia in the emergency room.

Methods

We identified 41,371 patients who presented to the ER of LDS Hospital, Salt Lake City, during a 25 months period (June 1995-June 1997). The discharge ICD-9 code (480-486) was the initial inclusion criterion for patients with CAP. There were 553 pneumonia patients during the 25-month period.

Fig. 1 illustrates the inclusion and exclusion criteria. Adapting our guideline criteria, we excluded 779 patients who were younger than 18 years. We excluded 4,540 patients who had a coded chief complaint that has been removed from the currently used list of codable chief complaints. Because the free text charting of the chief complaint has been reduced from 20% to 3%, we excluded 3,390 patients with a free text chief complaint that we were unable to code unambiguously. We excluded a total of 8,709 patients.

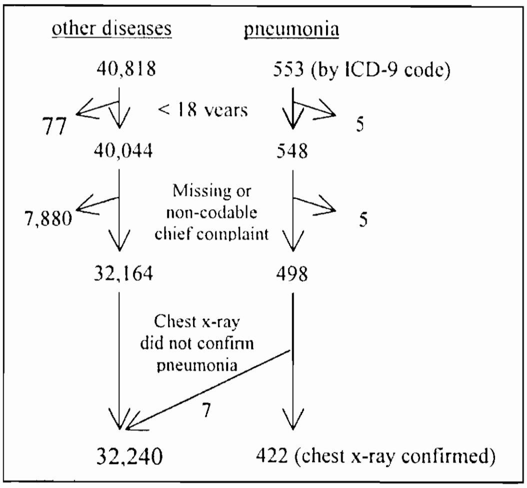

other diseases
pneumonia
40,818
553 (by ICD-9 code)
77
< 18 years
5
40,044
548
7,880
Missing or non-codable chief complaint
5
32,164
498
Chest x-ray did not confirm pneumonia
7
32,240
422 (chest x-ray confirmed)

Figure 1: inclusion and exclusion criteria of patients with and without CAP

For each of the remaining 32,662 patients, we extracted a total of 65 variables from the HELP System[9], of which 59 were coded and 6 free text. Only the first incidence of the data elements were considered. These data elements originated from different sources. A triage nurse captured the chief complaint, the current and past history, the current medication, allergies, and the vital signs. The nurse who took care of the patient during the encounter entered the patient's assessment. Lab values entered the HELP system through a laboratory interface. Free text information such as the current or past history was parsed for keywords. For all patients we calculated a risk factor similar to the algorithm used in our practice guideline.

The chest x-ray reports were extracted for all patients who had one or more chest x-rays taken within the first 72 hours of their encounter in the ER. The chest x-ray interpretation of the ER physician was generally available at a time when relevant decisions were made. However, the chest x-ray interpretation of the radiologist was considered the gold standard for diagnosing CAP. For all the patients with an ICD-9 code of CAP, we extracted both the dictated radiologist's chest x-ray reports and the ER physician's dictated clinical reports. We manually reviewed the reports for all the 498 patients with CAP. We applied the gold standard criteria and only included the 422 patients who had chest x-ray confirmed CAP. For the 32,163 patients without CAP, we identified 8,102 patients who had at least one chest x-ray taken within the 72-hour period. Their chest x-ray reports were parsed for keywords that were suggestive of CAP (e.g. "infiltrate", "consolidation", "no evidence of"). Based on a conservative algorithm we identified 995 patients who actually had other diseases than CAP, but whose dictated chest x-ray reports were compatible with CAP.

We developed the BN with Netica™, a software that performs Bayesian parameter learning.[10] A 300 MHz PC with 64 MB RAM was used for training and testing. We randomly assigned each of the 32,662 patients to one of three different subsets. We tested the BN with each of the three subsets while the two remaining subsets represented the training set.

An evaluation of the accuracy and the performance of the BN was determined applying measures that are typically used for clinical tests. These included the sensitivity, the specificity, and the positive predictive value.[11] The sensitivity and the specificity are important descriptive characteristics of a diagnostic test. To the clinician, however, the predictive value has a more clinically oriented meaning. In a patient with a positive test, it indicates, how many times a true or false positive result can be expected. Unfortunately, the prevalence of a disease influences the predictive value, whereas the sensitivity and the specificity are more consistent in the face of varying prevalence.

We calculated the receiver operating characteristic (ROC) curve to refine and evaluate different versions of the BN. The ROC curve is a graphic measure that plots corresponding pairs of

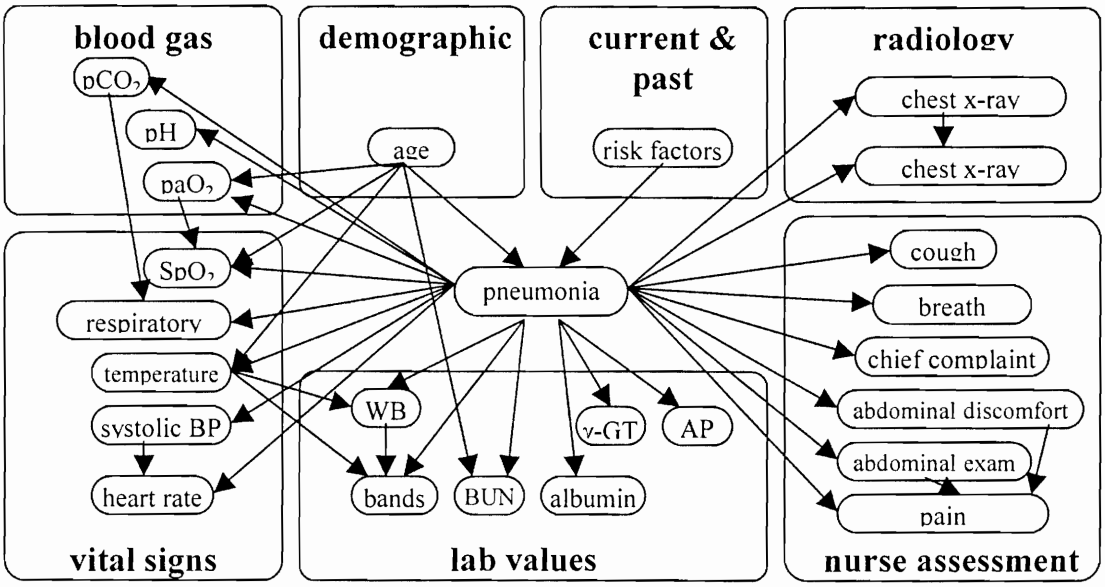


Figure 2: Structure of the Bayesian network for diagnosing CAP. All variables are available in the HELP system during a patient's encounter in the emergency room with the exception of the chest x-ray information ("chest x-ray positive").

the true positive rates (sensitivity) and the false positive rates (1-specificity).[12] The area under the ROC curve is a standard measure indicating the overall performance of a diagnostic test.[13] A lack of discriminatory ability exists when the sensitivity equals the specificity in which case the ROC curve is a 45° line and the corresponding area under the curve equals 0.5. Perfect discrimination exists when the sensitivity and the specificity equal 100% which yields an area under the ROC curve of 1.0.

Results

We implemented and evaluated more than 50 different network structures. The number of nodes in the different network structures ranged from 20 to 77 nodes. The size of the BN ranged from 262 kB to 8.6 MB and the run time for 100 cases ranged from 6 to 46 seconds.

The most parsimonious and most accurate BN contained 25 nodes, 38 links, and 10,100 conditional probabilities (Fig. 2). There were 3 dichotomous, 6 categorical and 16 continuously valued nodes. The node "chief complaint" contained 60 different states. The BN was 262 kB large and required 6 seconds to compute a probability of CAP for 100 cases.

The results of the three different test subsets are presented in Fig. 3. When the sensitivity was fixed at 95%, the corresponding specificity averaged 96.5%. The mean predictive value positive was 26.8% and the average area under the ROC curves of the three subsets was 0.9825.

For subset 3 we show the ROC curve (Fig 4), the 2x2 table (Fig. 5), and the most frequent discharge diagnosis (ICD-9) for the false positive group (Fig. 6.).

set	specificity (sensitivity fixed at 95%)	positive predictive value	area under the ROC curve
1	97.3 %	30.1 %	0.991
2	95.6 %	21.2 %	0.977
3	96.6 %	29.1 %	0.979

Figure 3: Results for the three testing set.

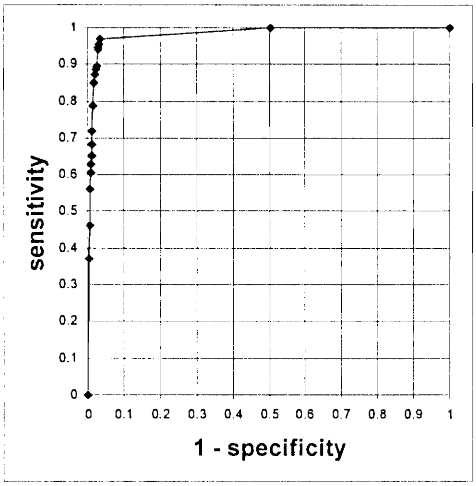


Figure 4: The ROC curve for test set 3. The area under the curve is 0.979

	patient with CAP	patients without CAP	total
BN positive	155	378	533
BN negative	8	10,622	10,630
total	163	11,000	11,163

Figure 5: 2x2 table of subset 3. The sensitivity is 95%, the specficity is 96.6%, and the predictive value positive is 29.1%.

congestive heart failure	32
aspiration pneumonia	19
urinary tract infection	17
fever of unknown origin	17
acute bronchitis or bronchiolitis	10
acute upper respiratory infections	9
other symptoms involving respiratory tract	8
status asthmatics	8
unspecified viral infections	6
pulmonary embolism	6
painful respiration	6
chest pain	6
acute respiratory distress	6
respiratory failure	5
pulmonary embolism	5
chronic obstructive asthma	5
chronic bronchitis with acute exacerbation	5
asthma unspecified	5
acute pyelonephritis	5

Figure 6: Discharge diagnosis (ICD-9) of false positive patients in subset 3. Only diseases with more than four patients are listed. They account for 47.6% of all the 378 false positive cases in this subset.

Discussion

As part of the feasibility study we have developed a BN for diagnosing community-acquired pneumonia in patients who present at the ER. The BN is being designed to screen every patient who presents to ER and to alert the ER physician about the possible presence of a patient with CAP. If the ER physician acknowledges the alert and confirms the diagnosis, the BN may trigger the computerized practice guideline for the management of the patient.

In our opinion three important requirements are important when the feasibility of a probabilistic real-time decision support system is evaluated. First, a sensitivity of 95% or higher combined with a very high specificity is mandatory. Second, most data elements in the BN and the updated probability have to be available while the patient is in the ER and before the physician has made his final decisions. Third, any additional data entry beyond the normal charting practices of the nurse or the physician should be eliminated or kept to an absolute minimum.

Considering the clinical application of the BN as a screening and alerting tool, the predictive value positive requires special attention. Although the combination of a 95% sensitivity with a 96.5% specificity is excellent, the predictive value is clinically more informative for the ER physicians. The predictive value specifies how many times the BN will alert the physician in patients with and without CAP. Our BN averages a predictive value positive of 26.8% which indicates that out of 4 issued alerts, only one would actually be a patient with CAP. The three false positive alerts, however, represent in large valid differential diagnosis to CAP (Fig. 6). Considering that the ER physicians see about 55 patients every day and one patient with CAP about every second day, the BN would alert them twice a day. Improving the predictive value will be difficult given the current level of specificity. An area under the ROC curve of 0.98 demonstrates that the overall accuracy and the discriminatory ability are exceptionally high.

To achieve this high level of accuracy, the clinical variables must be available during the patient's encounter in the ER. Clearly, an alert assists the physician only while the patient is in the ER. During the training and testing phase the BN has been presented at one single point with all the available data elements of a patient. The clinical work flow, however, is much different. The data are not present at one single point, but are gathered and charted over the entire time period that a patient is in the ER. In a BN, however, not all of the data have to be present. The BN accounts for the presence of uncertainty and can operate with varying amounts of missing data. Whenever the clinical information system records a new piece of data, the BN can incorporate the new evidence and update the joint probability. Experience suggests that the physician's compliance with a computerized decision support system depends on the amount of additional data elements that have to be entered. Currently, the BN incorporates variables that are part of the nurses' charting practices or originate from the laboratory. From both the nurse and the physician the BN does not need additional information, with the chest x-ray being the only exception.

The radiologists' chest x-ray interpretation is currently not available in the clinical information system during the patient's encounter in the ER. However, the chest x-ray is important for the diagnosis of CAP. For the following pilot study, we will need to prompt the ER physician to indicate whether the chest x-ray is positive or negative, a task which they have agreed to do. Since the HELP System records the time when a chest x-ray has been ordered, the BN will recognize for which patients it should prompt the ER physician. Although the radiologists' interpretation would be preferred, the current work flow does not provide a feasible procedure to include their interpretation while the patient is in the ER.

There are limitations in our study. First, the BN was designed for CAP in the ER of a tertiary care hospital. However, most of the patients with CAP are treated by their primary care physician. The representativeness of our database is therefore limited as our population represents a selected group of patients only. Patients that enter our ER may be more seriously ill and have a unrepresentative spectrum of causative organisms. Second, it is difficult to predict which data elements are gathered in a specific patient, and in which order they enter the HELP system. For instance, a blood gas sample is not obtained in every patient, and if it were, the moment will greatly vary at which the results will become available to the BN. With incomplete evidence the BN may cross the threshold and generate an alert, but may then drop the probability after more evidence becomes available. Defining a minimal set of instantiated variables may help to prevent premature alerts. Third, the database contains retrospective data, but a prospective evaluation in our ER is required to demonstrate whether the BN will perform as expected.

Conclusion

The results of our study have general significance for the application of Bayesian networks in a clinical environment. Clinical information systems are an accurate source for the assessment of probabilities that are required for the development of a probabilistic decision support system. The results obtained are encouraging and suggest that a Bayesian network may provide a promising method as a real-time decision support system in a clinical environment. We feel confident to perform a pilot study in the ER and test whether the Bayesian network may be an accurate component within a larger decision support system that assists emergency room physicians in the management of community-acquired pneumonia.

Acknowledgments

The author is supported with a grant from the Swiss National Science Foundation (31-44406.95). I thank my advisors, Reed M. Gardner and R. Scott Evans, for suggestions on the manuscript.

References

1. Garibaldi RA. Epidemiology of community-acquired respiratory tract infections in adults: incidence, etiology, and impact. *Am J Med*, 1985; 78:32-37.
2. Medicare and Medicaid statistical supplement. *Health Care Financ Rev*, 1995; 16 (September).
3. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med*,1997; 336: 243-50.
4. Dean NC, Bateman KA, Hadlock CJ, McKinstry CA, James BC.Use of Care Process Model for Community Acquired Pneumonia in Three Rural Counties. American Thoracic Society, Proceedings of the 1997 International Conference, San Francisco, CA.
5. Jensen FV. An introduction to Bayesian Networks, Springer-Verlag New York Inc., New York, 1997
6. Pearl J. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, California, 1988
7. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, 1988: Series B 50; 157-224.
8. Cooper GF. The Computational Complexity of Probabilistic Inference Using Belief Networks. Artificial Intelligence, 1990; 42:393-405
9. Kuperman GJ, Gardner RM, Pryor TA: HELP: A Dynamic Hospital Information System. New York: Springer-Verlag, 1991.
10. Netica, Application for Belief Networks and Influence Diagrams, User's Guide, Norsys® Software Corp., 1997.
11. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical Decision Making. Butterworth-Heinemann, Newton MA, 1988.
12. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear medicine*, 1978;8:283-98.
13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*,1982;143:29-36

CHAPTER 3

AN INTEGRATED DECISION SUPPORT SYSTEM FOR DIAGNOSING AND MANAGING PATIENTS WITH COMMUNITY-ACQUIRED PNEUMONIA

Dominik Aronsky, MD, Peter J Haug, MD

Proc AMIA Symp. 1999;:197-201.

(reprinted with permission from Hanley & Belfus, Inc.)

An Integrated Decision Support System for Diagnosing and Managing Patients with Community-Acquired Pneumonia

Dominik Aronsky, MD and Peter J. Haug, MD
Dept. of Medical Informatics, LDS Hospital/University of Utah, Salt Lake City, Utah

Decision support systems that integrate guidelines have become popular applications to reduce variation and deliver cost-effective care. However, adverse characteristics of decision support systems, such as additional and time-consuming data entry or manually identifying eligible patients, result in a "behavioral bottleneck" that prevents decision support systems to become part of the clinical routine. This paper describes the design and the implementation of an integrated decision support system that explores a novel approach for bypassing the behavioral bottleneck. The real-time decision support system does not require health care providers to enter additional data and consists of a diagnostic and a management component.

INTRODUCTION

In today's health care system, the computerized patient record (CPR) becomes the standard to collect, store, and report patient-related data. CPRs represent rich data sources that can be used to develop decision support systems (DSS) to support health care providers in caring for patients, decreasing variation and delivering cost-effective care.[1] Rule-based DSSs are frequently used to computerize guidelines of low or moderate complexity. Integrating more complex guidelines into a DSS that is intended for both disease- and patient-specific care, is highly desirable. However, the implementation and complete integration of disease- and patient-specific DSSs is challenging, because developers face both technical and behavioral problems that are difficult to overcome. Reasons that prevent a successful implementation are:

- The prevalence of the guideline-applicable disease in the population of interest is low, e.g., the prevalence of many diseases in an emergency department (ED) is below 2%, which often prevents the guideline from becoming part of daily practice.

- Required data elements are not available in a timely fashion, e.g., the radiologist's chest x-ray interpretation of a patient with suspected pneumonia usually becomes available at a time when treatment and admission decisions have already been made.

- The required detail of data to drive a guideline is not present, e.g., the presence of coexisting diseases is concealed in the free text portion of the patient's past history.[2]

- Data are not collected and represented in an easily computable or decidable format, e.g., the patient's present history or current medication is recorded in free text.[3]

- Additional data gathering in an appropriate format may be necessary, but is time-consuming and often results in double-charting[3], e.g., ECG measurements have to be entered explicitly, but are already available from the ECG tool.

- The lack of CPRs to automatically detect patients with certain diseases requires health care providers to manually identify patients, initiate and complete the guideline. However, the time involved in identifying a patient to the CPR and completing the guideline is often not available in a busy clinic.

The tasks involved in identifying eligible patients, initiating the computerized guideline and entering additional, patient-specific data represent a critical "behavioral bottleneck". In an attempt to master the behavioral bottleneck we have developed an integrated, real-time DSS that consists of a diagnostic and a management component. The diagnostic component is based on a Bayesian network and automatically identifies patients with CAP. The management component implements different guidelines for pneumonia patients. Both parts operate without requiring health care providers to enter additional data.

BACKGROUND

Decision support systems

The development and implementation of DSSs has proliferated in recent years.[4,5] Several applications have shown that DSSs are a suitable method to support health care providers. Some of them have demonstrated a positive impact on the delivery of appropriate and cost-effective patient care.[5,6] The integration of DSSs into a CPR has been predominantly accomplished with rule-based applications. Rule-based systems, however, are not able to represent the uncertainty and the variation inherent to the medical domain. A variety of methods that explicitly model medical uncertainty have been successfully applied and implemented particularly for diagnostic problems. Examples of methods include uncertainty factors, Fuzzy set logic, heuristic systems, neural networks, logistic regression, and Bayesian systems.[7] Although several methods have demonstrated impressive results in diagnostic performance, none has been completely integrated with a CPR. The lack of integration often prevents the use of DSS applications to support health care providers in real-time and at the point of care. In addition, diagnostic systems have not been explored for the purpose of detecting patients who have a particular disease of interest and who are eligible for guidelines.

Guidelines for community-acquired pneumonia

Several guidelines for CAP have been developed.[8-10] However, all of them are paper-based and complex jeopardizing implementation and dissemination. To promote the dissemination of the Pneumonia Severity Index (PSI), the original version, a logistic regression,[8] has been converted into a simpler scoring algorithm.[9] The PSI computes a severity risk class for patients with CAP. The algorithm evaluates 20 variables that are routinely available in the CPR at LDS Hospital (HELP System[11]) during the patient's initial encounter in the ED. All the data elements of the PSI are available in the HELP System, which increases the chances for automation. However, the ED physician still needs to interact with the HELP System and manually identify patients with CAP. Although this seems to be a simple task, many guideline implementations have failed because the importance of the "behavioral bottleneck" was underestimated.

A Bayesian network for patients with CAP

In an effort to tackle the underlying problems of the behavioral bottleneck, we have developed a Bayesian network (BN) that detects patients with CAP in an ED population.[12] BNs are graphical representations that are based on probability theory.[13] The probabilistic characteristics of BNs favor their the medical field. However, BNs have not been application for diagnostic purposes, particularly in integrated into clinical information systems and still need to demonstrate their value and applicability in an integrated clinical environment.

DESIGN AND IMPLEMENTATION

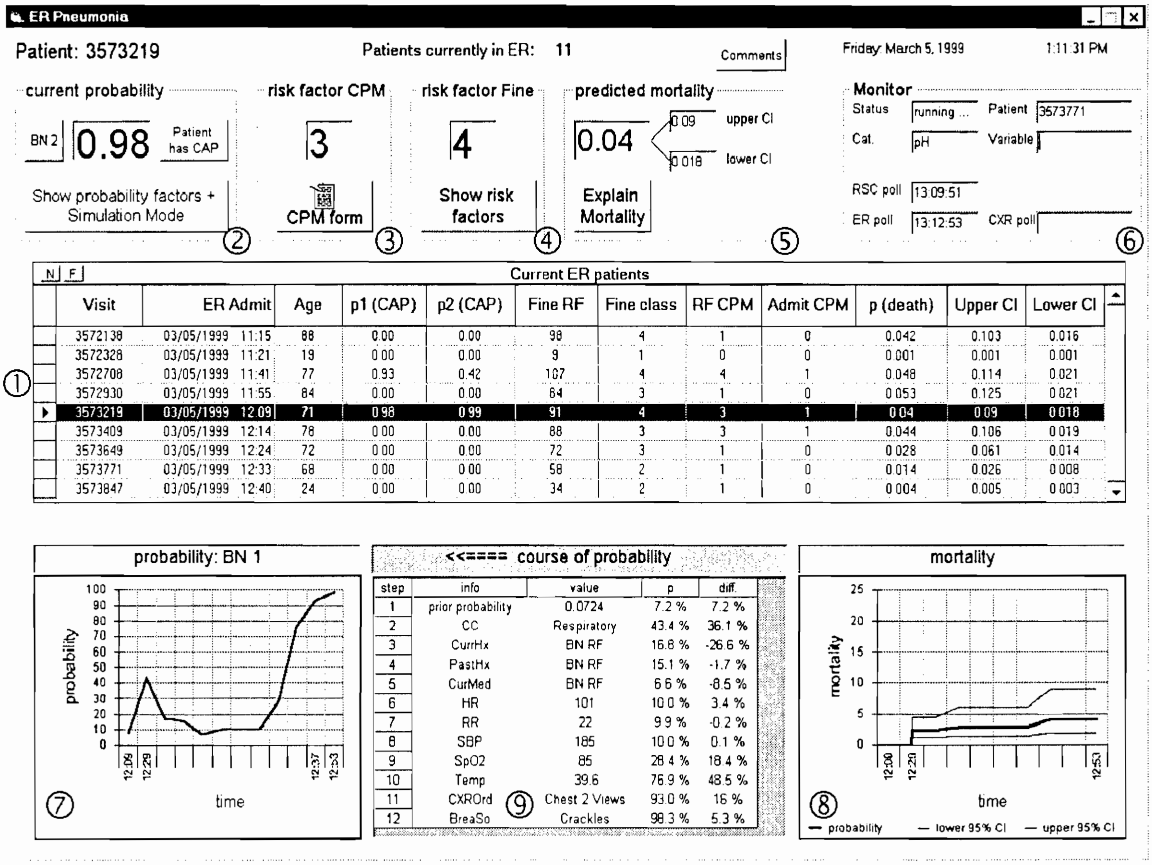
We have developed and implemented an integrated DSS that allows the identification of patients with CAP in an ED population and evaluates guidelines for the management of pneumonia patients. The purpose of the decision support application is to automatically provide the ED physicians with information that assists in the medical decision making process. Because the DSS uses only data elements that are routinely collected and available during a patient's encounter in the ED, the application does not require any additional data entry from the ED staff. In parallel to the medical decision making process, the DSS collects and evaluates data over the entire encounter of the patient. As soon as new patient data become available, the DSS combines the new evidence with already retrieved information. This real-time process avoids inconvenient interruption of the ED physician's work flow and allows them to review current information at any time during the patient's encounter.

The decision support system consists of two components: The first, population-based component identifies patients with CAP by applying a BN. The BN collects data and calculates a probability of CAP for all patients presenting to the ED. The second component provides information for the management of CAP patients by evaluating the two different versions of the PSI (logistic regression and scoring algorithm). In addition, a locally developed pneumonia guideline has been integrated into the DSS to respect the preferences of the local ED physicians.[10]

Data retrieval

Whenever a new patient is registered in the ED, the DSS is notified by the HELP System and adds the new patient to the list of current patients. For all the ED patients on the current list the DSS polls the HELP System for new data every 5-10 minutes. On every patient in the ED, the DSS retrieves up to 42 data elements that are required to calculate the probability of CAP and to evaluate the guidelines.

The data elements used for the DSS are demographic information, triage data, nurse assessment, results from laboratory tests and blood gas analyses, and chest x-ray related information. Demographic information include the patient's age and sex. Triage data consist of the chief complaint, the present and past medical history, the current medication, allergies, and initial vital signs (heart rate, respiratory rate, systolic blood pressure, oxygen saturation, and temperature). The ED nurse assessment is charted directly into the HELP System. The DSS retrieves information about breath sounds, cough, abdominal exam, abdominal discomfort, pain



ER Pneumonia

Patient: 3573219

Patients currently in ER: 11

Friday March 5, 1999 1:11:31 PM

current probability: BN 2 0.98 Patient has CAP

risk factor CPM: 3

risk factor Fine: 4

predicted mortality: 0.04 (upper CI 0.09, lower CI 0.018)

Monitor: Status running ... Patient 3573771 Cat. pH Variable

RSC poll 13:09:51

ER poll 13:12:53 CXR poll

Current ER patients

	Visit	ER Admit	Age	p1 (CAP)	p2 (CAP)	Fine RF	Fine class	RF CPM	Admit CPM	p (death)	Upper CI	Lower CI
	3572138	03/05/1999 11:15	88	0.00	0.00	98	4	1	0	0.042	0.103	0.016
	3572328	03/05/1999 11:21	19	0.00	0.00	9	1	0	0	0.001	0.001	0.001
	3572708	03/05/1999 11:41	77	0.93	0.42	107	4	4	1	0.048	0.114	0.021
	3572930	03/05/1999 11:55	84	0.00	0.00	84	3	1	0	0.053	0.125	0.021
▶	3573219	03/05/1999 12:09	71	0.98	0.99	91	4	3	1	0.04	0.09	0.018
	3573409	03/05/1999 12:14	78	0.00	0.00	88	3	3	1	0.044	0.106	0.019
	3573649	03/05/1999 12:24	72	0.00	0.00	72	3	1	0	0.028	0.061	0.014
	3573771	03/05/1999 12:33	68	0.00	0.00	58	2	1	0	0.014	0.026	0.008
	3573847	03/05/1999 12:40	24	0.00	0.00	34	2	1	0	0.004	0.005	0.003

<<==== course of probability

step	info	value	p	diff.
1	prior probability	0.0724	7.2 %	7.2 %
2	CC	Respiratory	43.4 %	36.1 %
3	CurrHx	BN RF	16.8 %	-26.6 %
4	PastHx	BN RF	15.1 %	-1.7 %
5	CurMed	BN RF	6.6 %	-8.5 %
6	HR	101	10.0 %	3.4 %
7	RR	22	9.9 %	-0.2 %
8	SBP	185	10.0 %	0.1 %
9	SpO2	85	28.4 %	18.4 %
10	Temp	39.6	76.9 %	48.5 %
11	CXROrd	Chest 2 Views	93.0 %	16 %
12	BreaSo	Crackles	98.3 %	5.3 %

Fig. 1: The main screen of the decision support system (supervisory mode). The patient highlighted in the ED patient table has a high probability of pneumonia according to the Bayesian network. The summary data are displayed in the upper section. The charts in the lower section show the probability changes after the patient has been in the ED for an hour. The left chart graphs the probability changes over time for pneumonia and the right chart plots the changes for the associated mortality. In the table between the graphs the user can review the data elements that have become available so far and that have been involved in the calculation of the probabilities. The circled numbers correspond to the numbers in the text.

characteristics, and mental status. Laboratory values include sodium, blood urea nitrogen, creatinine, albumin, glucose, white blood count, bands, hematocrit, hemoglobin, and the results of liver function tests. If a blood gas analysis is performed, the pH, the paO2, and the pCO2 are retrieved. If a series of values are present in a patient, the DSS considers only the initial value. Once the value has been obtained, the DSS does not poll for the same data element again. Encounter related data include the time of admission and discharge, time of chart completion, and final discharge disposition.

If a chest x-ray is ordered on a patient, the DSS retrieves both the time when the order has been placed and the time when the chest x-ray has been completed. The ED physicians review the chest x-ray on a radiology viewing station in the ED and usually enter a short, often abbreviated interpretation of their finding into a dedicated comment field. If the chest x-ray of a patient supports the diagnosis of CAP they use “p+” as an abbreviation and “e+” to indicate the presence of an effusion.

Data evaluation

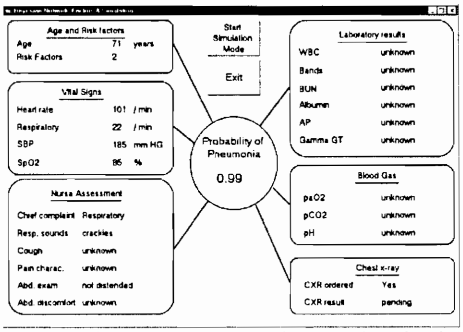
Whenever new information becomes available, the BN incorporates the new data element and updates the probability of pneumonia. Similarly, the pneumonia guidelines are updated. The user can change, correct or add values in either component of the decision support system. If the user modifies data elements, the BN and the guidelines are updated immediately. User changes in the models are captured in a log file.

Accessing and presenting data

The DSS displays the principal information on the main screen (Fig. 1). The main screen is organized into three major sections and the displayed data depend on whether the DSS is in supervisory or user mode.

The table in the center of the main screen (①; the numbers in the text refer to Fig. 1) lists all the patients that are currently in the ED together with the current probability of CAP and the summary data of the evaluated guidelines. If a patient is highlighted in the table of current ED patients, more detailed information is available in the upper and the lower sections.

The upper section of the main screen allows the user to investigate the probability of CAP and the results of the guidelines in detail. The section with the current probability (②) has a button labeled "Show probability factors + simulation mode"). This allows the user to review the values and the data elements that are involved in calculating the probability of CAP (Fig. 2). Additionally, users can initiate a simulation session where they can change the values of the BN and investigate the impact on the probability. The next section (③) displays the summary risk factor for the locally developed guideline. The complete guideline can be called with the button "CPM form". The next two sections display the results of the two PSI versions, namely the risk factor (④) from the scoring algorithm and the probability estimate from the logistic regression (⑤). The PSI scoring algorithm (Fig. 3) and the data elements involved in the logistic regression can be accessed and further investigated with the respective buttons. The monitor frame (⑥) allows the user to view the activity of the interfaces to the clinical information system and the radiology database. The lower section of the main screen graphically displays how the probability of CAP (⑦) and the probability of death (original PSI version) (⑧) have developed over time. The user has the ability to obtain an explanation in the table between the two charts by selecting a point of interest from either graph. The table (⑨) lists the data elements that are currently available from the selected patient, the resulting absolute probability, and the difference of probability caused by incorporating the respective data element.



Age and Risk factors

Age	71	years
Risk Factors	2	

Vital Signs

Heart rate	101	/ min
Respiratory	22	/ min
SBP	185	mm HG
SpO2	86	%

Nurse Assessment

Chief complaint	Respiratory
Resp. sounds	crackles
Cough	unknown
Pain charac.	unknown
Abd. exam	not distended
Abd. discomfort	unknown

Laboratory results

WBC	unknown
Bands	unknown
BUN	unknown
Albumin	unknown
AP	unknown
Gamma GT	unknown

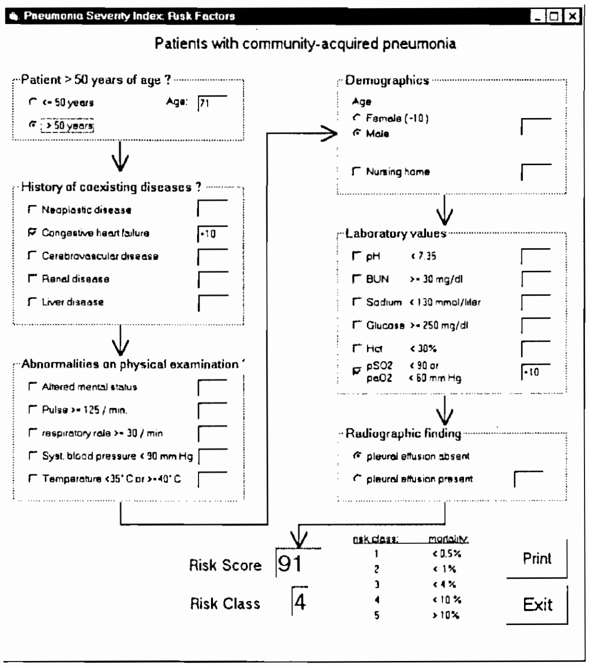
Blood Gas

paO2	unknown
pCO2	unknown
pH	unknown

Chest x-ray

CXR ordered	Yes
CXR result	pending

Fig. 2: The data elements that are involved in computing the probability are shown for the selected patient from the main screen (Fig. 1). Note that the chest x-ray has been ordered before the laboratory values are known. The user can start the simulation mode and explore the BN by changing values.



Patients with community-acquired pneumonia

Patient > 50 years of age ?
○ <= 50 years Age: 71
● > 50 years

History of coexisting diseases ?
☐ Neoplastic disease
☑ Congestive heart failure +10
☐ Cerebrovascular disease
☐ Renal disease
☐ Liver disease

Abnormalities on physical examination ?
☐ Altered mental status
☐ Pulse >= 125 / min.
☐ respiratory rate >= 30 / min
☐ Syst. blood pressure < 90 mm Hg
☐ Temperature <35° C or >=40° C

Demographics
Age
○ Female (-10)
● Male
☐ Nursing home

Laboratory values
☐ pH < 7.35
☐ BUN >= 30 mg/dl
☐ Sodium < 130 mmol/liter
☐ Glucose >= 250 mg/dl
☐ Hct < 30%
☑ pSO2 < 90 or paO2 < 60 mm Hg +10

Radiographic finding
● pleural effusion absent
○ pleural effusion present

Risk Score 91

Risk Class 4

risk class:	mortality:
1	< 0.5%
2	< 1%
3	< 4%
4	< 10%
5	> 10%

Fig. 3: The PSI scoring algorithm is shown for the selected patient from the main screen (Fig. 1). The DSS evaluates the PSI scoring algorithm with the values that are currently available. The pathological values are marked, the attributable points entered and summed. The resulting risk class is derived from the risk score. The user can change, add, or delete values. Any changes by the user triggers an update of the BN and the guidelines.

Environment

Access to the decision support system is password-protected. The DSS is a client-server application and is implemented in the VisualBasic programming language. The DSS resides on a client PC and uses MS Access for the data management. The Bayesian network inference engine is accessed through an API (application programming interface). The connection to the radiology database is established through ODBC (open database connectivity). The DSS communicates to the HELP System through an interface that was implemented using XML (Extended Markup Language).

DISCUSSION

Our real-time DSS explores a new approach to support busy clinicians with guideline information. The diagnostic component of the DSS automatically detects patients with CAP from a broad and unrestricted ED population. Without requiring additional data entry, the diagnostic probability of CAP as well as the guideline information are available in real time. Time-consuming interaction between the user and the DSS are eliminated because the DSS works only with data elements that are routinely available in the CPR. The DSS provides additional and presumably useful information that clinicians can integrate into the decision making process.

The underlying design of our DSS addresses particularly the behavioral bottleneck. Clinicians are reluctant to use computerized guidelines that require additional data entry and consume considerable time and effort to be completed. Additionally, many diseases for which guidelines exist have a low prevalence. The combination of these adverse characteristics often prevent guidelines to become part of the clinical routine. Many guideline implementations, paper-based or computerized, have failed because they were not able to bypass the behavioral bottleneck. If a DSS is able to automatically detect patients with certain diseases, the behavioral bottleneck may be mastered and the computerized implementation of guidelines becomes more feasible.

However, the entire DSS depends primarily on the diagnostic accuracy of the BN. Although the BN has performed well on historical data[12], we are assessing the diagnostic performance in a two-phase prospective study. In the first phase the diagnosis of the ED physicians and the BN are compared separately against a gold standard. The first phase will allow us to quantify the diagnostic accuracy of the ED physicians and the BN separately. However, this phase does not reflect the intended clinical situation, but will result in baseline data. The second phase will reflect the actual intended clinical situation where the information of the DSS will be available to clinicians. This phase will evaluate the supplemental information from the DSS on the physician's decision making.

If the diagnostic component proves to reliably detect patients with CAP, the chosen approach can be used to mark the records of pneumonia patients in the CPR. Once the CPR contains a marker for patients with CAP, additional guidelines can be initiated. For example, known pneumonia patients can be evaluated with respect to vaccination guidelines, to ICU admission criteria, and to hospital discharge criteria.

Acknowledgments

This work has been supported by the Swiss National Science Foundation Grant 31-44406.95.

References

1. Berner ES. Clinical Decision Support Systems: Theory and Practice. New York: Springer-Verlag, 1999.
2. Tierney WM, Overhage JM, McDonald CJ. Toward electronic medical records that improve care. Ann Intern Med 1995;122:725-6.
3. Barnett GO, Winickoff RN. Quality assurance and computer-based patient records. Am J Public Health 1990;80:527-8.
4. Miller RA. Medical diagnostic decision support systems – past, present, and future: A threaded bibliography and commentary. J Am Med Inform Assoc. 1994 Jan-Feb;1(1):8-27.
5. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA 1998 Oct 21;280(15):1339-46.
6. Evans RS, Pestotnik SL, Classen DC et al. A computer-assisted management program for antibiotics and other antiinfective agents. N Engl J Med 1998 Jan 22;338(4):232-8.
7. Russell SJ and Norvig P. Artificial Intelligence: A Modern Approach. Prentice Hall, New Jersey, 1995.
8. Fine MJ, Hanusa BH, Lave JR, et al. Comparison of a disease-specific and a generic severity of illness measure for patients with community-acquired pneumonia. J Gen Intern Med. 1995;10:359-68.
9. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med 1997;336:243-50.
10. Dean NC, Bateman KA, Hadlock CJ, McKinstry CA, James BC. Use of Care Process Model for Community Acquired Pneumonia in Three Rural Counties. American Thoracic Society, Proceedings of the 1997 International Conference, San Francisco, CA.
11. Kuperman GJ, Gardner RM, Pryor TA: HELP: A Dynamic Hospital Information System. New York: Springer-Verlag, 1991.
12. Aronsky D, Haug PJ. Diagnosing Community-Acquired Pneumonia with a Bayesian Network. Proc Amia Symp. 1998;:632-6.
13. Jensen FV. An introduction to Bayesian Networks, Springer-Verlag New York Inc., New York, 1997.

CHAPTER 4

ASSESSING THE QUALITY OF CLINICAL DATA IN A COMPUTER-BASED RECORD FOR CALCULATING THE PNEUMONIA SEVERITY INDEX

Dominik Aronsky, MD, Peter J Haug, MD

J Am Med Inform Assoc. 2000 Jan-Feb;7(1):55-65

(reprinted with permission from Hanley & Belfus, Inc.)

Research Paper ■

Assessing the Quality of Clinical Data in a Computer-based Record for Calculating the Pneumonia Severity Index

DOMINIK ARONSKY, MD, PETER J. HAUG, MD

Abstract **Objective:** This study examined whether clinical data routinely available in a computerized patient record (CPR) can be used to drive a complex guideline that supports physicians in real time and at the point of care in assessing the risk of mortality for patients with community-acquired pneumonia.

Setting: Emergency department of a tertiary-care hospital.

Design: Retrospective analysis with medical chart review.

Patients: All 241 inpatients during a 17-month period (Jun 1995 to Nov 1996) who presented to the emergency department and had a primary discharge diagnosis of community-acquired pneumonia.

Methods/Main Outcome Measures: The 20 guideline variables were extracted from the CPR (HELP System) and the paper chart. The risk score and the risk class of the Pneumonia Severity Index were computed using data from the CPR alone and from a reference standard of all data available in the paper chart and the CPR at the time of the emergency department encounters. Availability and concordance were quantified to determine data quality. The type and cause of errors were analyzed depending on the source and format of the clinical variables.

Results: Of the 20 guideline variables, 12 variables were required to be present for every computer-charted emergency department patient, seven variables were required for selected patients only, and one variable was not typically available in the HELP System during a patient's encounter. The risk class was identical for 86.7 percent of the patients. The majority of patients with different risk classes were assigned too low a risk class. The risk scores were identical for 72.1 percent of the patients. The average availability was 0.99 for the data elements that were required to be present and 0.79 for the data elements that were not required to be present. The average concordance was 0.98 when all a patient's variables were taken into account. The cause of error was attributed to the nurse charting in 77 percent of the cases and to the computerized evaluation in 23 percent. The type of error originated from the free-text fields in 64 percent, from coded fields in 21 percent, from vital signs in 14 percent, and from laboratory results in 1 percent.

Conclusion: From a clinical perspective, the current level of data quality in the HELP System supports the automation and the prospective evaluation of the Pneumonia Severity Index as a computerized decision support tool.

■ JAMIA. 2000;7:55–65.

Affiliation of the authors: LDS Hospital, University of Utah, Salt Lake City, Utah.

The first author was supported in part by grant 31-44406.95 from the Swiss National Science Foundation.

Correspondence and reprints: Dominik Aronsky, MD, Department of Medical Informatics, LDS Hospital, 8th Avenue and C Street, Salt Lake City, UT 84143; e-mail: ⟨lddarons@ihc.com⟩.

Received for publication: 5/19/99; accepted for publication: 9/15/99.

Disseminating and implementing paper-based guidelines in everyday practice remains a major challenge.[1-4] A common reason for the reluctance to use the guidelines is the time required to complete them. Especially in hospitals with a computerized patient record (CPR), paper-based guidelines represent a duplication of data collection that should be avoided.[5] Computerizing a guideline is an attractive and effective means of avoiding the duplication of data collection and the time-consuming manual completion of guidelines.[6] Clinical information systems have the potential to drive guidelines[7,8] and minimize or eventually eliminate additional data collection from the health care providers. Few guideline-based decision support systems are integrated into existing CPRs, support clinicians in real time, and do not require additional data collection.[8-11] One reason for the sparsity of real-time computerized guidelines that do not require additional data entry is that clinically relevant data are not sufficient[12] or are not represented in an easily retrievable and computable format.[5]

A successful approach to computerizing guidelines is the capturing of essential guideline data on a structured encounter screen.[13,14] Guidelines with specifically designed encounter screens capture data in a computable format and have demonstrated both an improvement in documentation[14] and a positive effect on patient outcomes.[15] However, as the guideline increases in complexity, the time required to enter data grows. In contrast, the complexity of a guideline is often necessary to deliver recommendations that are patient-specific. An alternative and desirable approach to delivering patient-specific advice is to integrate a guideline into an existing CPR, taking advantage of the available data.[16,17] This approach may sacrifice data quality because the variables are not specifically collected for driving a guideline but rather are captured for documenting routine patient care.

Even if all the required data elements of a guideline exist in a CPR, the representation format affects the accuracy of guideline recommendations. Data quality needs to be assessed prior to implementation, because erroneous recommendations based on inaccurate data influence guideline acceptance and may raise liability issues.[18] Data quality is a fundamental issue when guidelines are integrated into a CPR. Only a few studies, however, focus on assessing the data quality or accuracy in a CPR.[19,20]

The objective of this study was to determine the data quality of variables routinely collected in an emergency department in driving the Pneumonia Severity Index (PSI).[21] The PSI guideline allows emergency department physicians to assess the risk of mortality in patients who have community-acquired pneumonia. Although the present form of the guideline assures patient-specific recommendations, the complexity remains too high to be easily memorized by clinicians. Computerizing and integrating the guideline into a CPR with verified levels of data quality is a desirable method to deliver real-time support.[22] Our goal was to test whether data routinely available during a patient's encounter in the emergency department can be used to evaluate the PSI and provide physicians with real-time decision support for the management of pneumonia patients at the point of care.

Methods

Following the methodological recommendations of Hogan and Wagner,[20] we describe the setting and charting process in detail.

Setting

LDS Hospital is a 520-bed tertiary-care and university teaching hospital in Salt Lake City, Utah. During the study period, the emergency department staff included 12 full-time, board-certified emergency department physicians. Because of personnel turnover, the nursing staff changed during the study period. There were, generally, 12 full-time and 30 part-time nurses. The emergency department staff cares for more than 25,000 patients per year and uses the HELP (Health Evaluation through Logical Processing) System for data recording and reporting.[23,24] In the emergency department there are 24 HELP terminals, the majority residing in patients' rooms.

The HELP System (version 15) is an inpatient CPR that has a long history and is well known for several integrated decision support systems.[25] It is a commercially available clinical information system (3M HIS, Murray, Utah) that runs on a mirrored Tandem mainframe computer with 12 central processing units. It was developed and is maintained in PAL (PTXT application language), which is a proprietary language of 3M, and in TAL (Tandem application language), which is a proprietary language of Compaq (previously Tandem, located in Houston, Texas). System downtime is 0.15 percent per year.

Defining either the computer-based or the paper-based chart as the official patient record is usually not feasible. Although the HELP System contains most a patient's data, this CPR is complemented with handwritten admission notes, progress notes, and additional forms that reside only in the paper-based record. Similarly, the HELP System contains information that does not enter the paper-based chart unless it is printed and inserted there on request. Thus, the official patient record (the Record) is the combined data from both the CPR and the paper-based chart.

Inclusion and Exclusion Criteria

We included all inpatients who were at least 18 years of age and had a primary discharge ICD-9-CM diagnosis of community-acquired pneumonia (ICD-9 codes 480.0–486.9) during a 17-month period (June 1995 to November 1996).

The chief complaint is a mandatory entry without which the emergency department nurse cannot chart in the CPR. For patients who have cardiac arrest or other life-threatening conditions, nurse charting is performed exclusively on paper and does not enter the CPR. Therefore, we identified patients who had cardiac arrest or another life-threatening condition by the absence of a chief complaint. These patients were excluded from the study. We also excluded patients with pneumonia who should have been admitted directly to the hospital but were admitted first to the emergency department because of a logistical misunderstanding. Because of the misunderstanding, these patients were not seen by an emergency department physician.

The PSI guideline excludes patients who have histories of acquired immunodeficiency syndrome, have a positive titer of HIV antibodies, have been transferred from another acute-care hospital, or have been hospitalized in the seven days prior to the current emergency department encounter. We included patients who met the PSI exclusion criteria, however, because the study did not include the assessment of data quality for variables that evaluate a patient's eligibility criteria for the guideline.

Data Collection in the Emergency Department

The emergency department nurses collect the majority of the data elements required for the PSI. When a patient presents to the emergency department, a registration clerk collects the patient's demographic information. A triage nurse collects the information about chief complaint, current and past histories, and current medication and measures vital signs, including temperature, blood pressure, and heart and respiratory rates. The triage nurse enters the patient's information directly into the HELP System. At the end of the triage, the nurse prints the captured information and attaches the form to the patient's chart. The form is then used by the emergency department physician to record additional findings.

After the patient is transferred to an emergency department room, an assigned nurse charts a more detailed assessment. The assessment is entered in coded form, but the nurse can also chart findings in free text. While the patient is being evaluated, the nurse may measure the vital signs again. Patients whose conditions are considered urgent do not present to the triage nurse but are admitted directly to an emergency department room, where the assigned nurse collects the triage information in addition to the assessment. After evaluating the patient, the emergency department physicians add information to the form printed by the triage nurse.

A clerk orders laboratory tests or radiology examinations through the HELP System. The laboratory results enter the HELP System through an interface with the laboratory computer. The radiologic images are reviewed by the physicians on a radiology workstation in the emergency department. The emergency department physicians review patient information on the HELP System. However, they are not involved in capturing data or entering orders. At the end of the patient's encounter, the nurse discharges the patient from the emergency department and finishes the charting process by recording the discharge time. The physician's dictation and the radiologist's x-ray interpretation enter the HELP System as free-text reports.

The Pneumonia Severity Index Guideline for Community-acquired Pneumonia

The PSI guideline is a severity scoring system that accesses the risk of mortality for pneumonia patients. The PSI was originally developed as a logistic regression[26] but was later converted to a scoring algorithm to ease clinical use and promote the dissemination of the guideline in different settings.[21] The PSI guideline is a two-step algorithm (Figure 1). The first step evaluates 11 variables from the physical examination and the patient's current and past histories. Patients who are 50 years old or younger and have no abnormal findings on the physical examination and the current and past histories are assigned to risk class 1. Otherwise, the patients are assigned to a higher risk class, which is determined in the second step.

The second step evaluates nine additional variables from laboratory tests and the chest x-ray (Table 1). To determine the appropriate risk class in the second step, the physician must first calculate a risk score. The risk score is the sum of points assigned to each of the 20 PSI variables. Finally, the risk class is derived from the risk score and corresponds to a patient's probability of dying (Table 2). The developers of the PSI have suggested that the risk class can be applied as an admission criteria.[21] Patients at low risk of dying might be managed as outpatients, whereas patients at high risk should be admitted to the hospital.

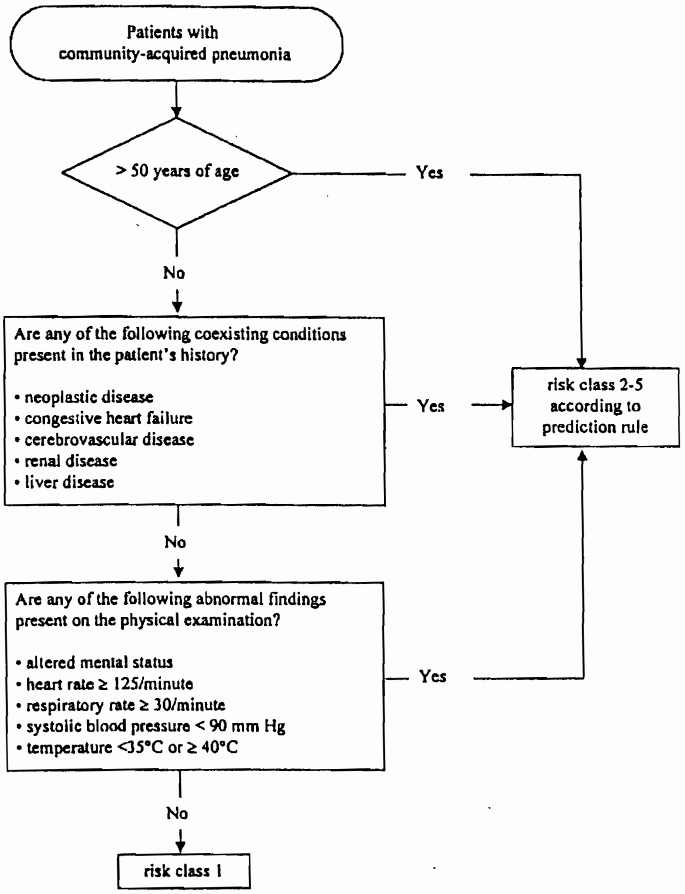

Patients with community-acquired pneumonia
> 50 years of age
Yes
No
Are any of the following coexisting conditions present in the patient's history?
• neoplastic disease
• congestive heart failure
• cerebrovascular disease
• renal disease
• liver disease
Yes
risk class 2-5 according to prediction rule
No
Are any of the following abnormal findings present on the physical examination?
• altered mental status
• heart rate ≥ 125/minute
• respiratory rate ≥ 30/minute
• systolic blood pressure < 90 mm Hg
• temperature <35°C or ≥ 40°C
Yes
No
risk class 1

Figure 1 First step of the Pneumonia Severity Index (PSI) scoring algorithm. On the basis of the patient's historical data and vital signs, the clinician assigns the patient to either risk class 1 or a higher risk class. For patients assigned to a higher risk class, a risk score is calculated and used by the clinician to determine the appropriate risk class. (Reprinted with permission from Fine et al.[21] © 1997, New England Journal of Medicine.)

Evaluating the Pneumonia Severity Index

For each patient we calculated the PSI risk score and risk class using data that were available in the CPR during the patient's encounter in the emergency department. The PSI risk score and risk class computed from the CPR data were then compared with the risk measures that were computed using data that originated in the patient's Record (the combined CPR and paper-based record). All information that was actually available in any format during the patient's encounter in the emergency department and that originated in the patient's Record represented our reference standard. The reference standard corresponds to the best information available while the patient was in the emergency department.

For all patients in the study we retrieved the PSI parameters from the CPR through database queries. To assess the five variables of the PSI that involve coexisting conditions (neoplastic disease, congestive heart failure, cerebrovascular disease, renal disease, and liver disease), we constructed a list of terms representing each coexisting condition. The list of terms was compiled by review of the free-text fields of the nurse-charting entries in the CPR for patients who were seen in the six months following the study period (Dec 1996 to May 1997). For patients in the study we inferred the presence of disease if one of the terms was present in the free-text field of the current history, the past history, or the current medication. We considered only patient information that was recorded while the patient was in the emergency department. The emergency department encounter started at the time a patient was registered by either the registration clerk or the emergency department nurse and ended at the time the patient was admitted to the hospital.

Table 1 ■

Variables Evaluated in the Second Step of the Pneumonia Severity Index (PSI) Scoring Algorithm

Patient Characteristic	Source of Data in the HELP System	Type of Data in the HELP System	Points Assigned for Abnormality
Demographics:			
Age:			
Men	Registration	Coded	Age (year)
Women	Registration	Coded	Age (year) −10
Nursing home	Triage	Free text	+10
Coexisting illnesses:			
Neoplastic disease	Triage	Free text	+30
Liver disease	Triage	Free text	+20
Congestive heart failure	Triage	Free text	+10
Cerebrovascular disease	Triage	Free text	+10
Renal disease	Triage	Free text	+10
Physical examination findings:			
Altered mental status	Nurse assessment	Coded	+20
Respiratory rate	Triage	Numeric	+20
Systolic blood pressure	Triage	Numeric	+20
Temperature	Triage	Numeric	+15
Heart rate	Triage	Numeric	+10
Laboratory findings:			
Arterial pH	Laboratory	Numeric	+30
Blood urea nitrogen	Laboratory	Numeric	+20
Sodium	Laboratory	Numeric	+20
Glucose	Laboratory	Numeric	+10
Hematocrit	Laboratory	Numeric	+10
p_aO_2 or SpO_2	Laboratory or triage	Numeric	+10
Radiographic finding:			
Pleural effusion	ED physician's report	Free text	+10

NOTE: The second step of the PSI scoring algorithm evaluates 20 variables to establish a risk score for the patient. In the HELP system the variables differ in format and have different sources. The score for the PSI is calculated by adding the patient's age and the points assigned for each abnormal finding. ED indicates emergency department.

Table 2 ■

Association of Risk Score with Risk Class and Mortality in the Pneumonia Severity Index (PSI) Scoring Algorithm

Risk Score	Risk Class	Mortality (%)
Based on step one	1	<0.5
≤70	2	0.5–0.9
71–90	3	1–3.9
91–130	4	4–10
>130	5	>10

NOTE: The second step of the PSI scoring algorithm assigns the patient to a risk class based on the calculated risk score. The assigned risk class corresponds to a probability of death.

For every patient in the study we obtained the paper-based charts from the medical records department. To abstract data from the patient's record, a self-coding data sheet was completed. The self-coding data sheet has been previously used in our institution for collecting data from the charts of pneumonia patients. As a safeguard against introducing bias into the unblinded review process, the records of 24 patients (10 percent) were randomly selected and were re-evaluated after a two-month interval.

Evidence of abnormal findings in the CPR or the Record that became available after the patient had left the emergency department was not considered for the evaluation of the PSI. For example, if a blood gas analysis was in process but the results were not available while the patient was present in the emergency department, the results were not included in the computation of the PSI. If a series of data elements from the same category was present, only the first measurement, and not the most abnormal measurement, was considered. For example, if the patient's initial systolic blood pressure was 100 mmHg, but subsequent measurements fell below the critical value of the PSI because the patient's condition worsened, only the initial value of 100 mmHg was included in calculating the risk score.

During the study period, the emergency department physician's interpretation of chest x-rays was not available, because information from the radiology database could not be accessed in real time. However, the emergency department physician's interpretation became available for real-time evaluation after the study period. A current implementation of the PSI would include the emergency department physician's interpretation of chest x-rays; consequently, we included the variable "pleural effusion" in the evaluation of the PSI. The principal reason for the emergency department physicians' real-time documentation of their chest x-ray interpretations is to facilitate fast and successful communication with the radiologists. To determine whether a pleural effusion was present, we manually retrieved the emergency department physician's chest x-ray findings from the emergency department physician's free-text report. Discrepancies between the emergency department physician's interpretation of the chest x-ray and that of the radiologist were not considered, because the radiologist's interpretation is not available during the patient's encounter.

Outcome Measures

The risk class is the clinically relevant measure that provides the emergency department physician with objective information about the patient's risk of mortality. However, for evaluating the impact of data quality of the CPR on the PSI, we determined the differences for both the risk scores and the risk classes, because an aberrant value might influence the patient's risk score without changing the risk class.

We assessed the data quality of the 20 PSI guideline variables along the two dimensions "availability" and "concordance." *Availability* is the proportion of observations from the reference standard that were actually recorded in the CPR. *Concordance* is the proportion of observations that are identically recorded in the CPR and the reference standard.

Because the emergency department created different standard sets of clinical variables that are required to be collected during a patient's encounter, we distinguished between the availabilities of required and optional PSI variables. The standard sets depend on the patient's chief complaint and triage category. The triage category of patients with suspected pneumonia is commonly "nonurgent." For nonurgent patients, the standard set of variables required for computing the PSI includes age, gender, the current and past history, the current medication, the systolic blood pressure, the heart and respiratory rate, and the temperature. The standard data set covers 12 PSI variables that are required to be present. The remaining eight PSI variables are not part of the standard data set and are, therefore, not required to be collected for every pneumonia patient. For example, a blood gas analysis does not represent an indispensable test for the diagnosis or management of pneumonia. Accordingly, "arterial pH" represents an optional PSI variable that is not required to be collected.

We quantified concordance for all variables of the PSI except "pleural effusion," because the data source for pleural effusion was identical for both the input variable and the reference standard. Concordance was assessed both when an individual patient was the unit of analysis and when all the variables were considered as one complete set. To examine the discordant data elements in more detail, we determined the types of error and causes of error. To explore the types of error, we stratified the errors into four categories, which depend on the data format and the source of the variables—free text, coded data, laboratory data, and vital signs. To examine the causes of error, we stratified the errors into two categories. One category included errors that were attributable to the emergency department charting process, and one category contained errors that originated from the computerized evaluation of the PSI algorithm. For example, if the emergency department nurse did not record that the patient had congestive heart failure or renal disease, the omission was categorized as an emergency department charting error. If the emergency department nurse misspelled a term, the error was categorized as a system error due to computerization of the PSI, because a parsing algorithm should be able to detect common misspellings. Distinguishing between different sources of errors means determining whether the errors are related to the actual computerization of the guideline or to the emergency department charting process in general. For example, if emergency department physicians do not document the presence of a pleural effusion, the resulting error is unrelated to the computerization of the guideline. Determining different sources of errors supports future efforts to improve the data quality for driving the PSI guideline.

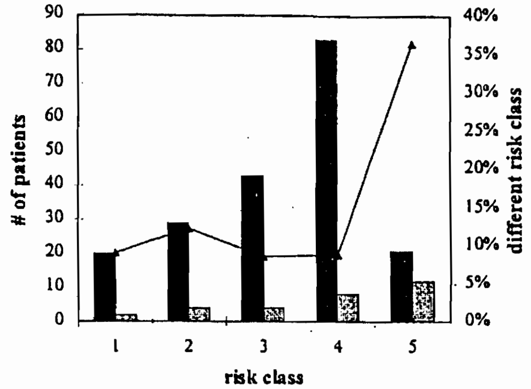


Figure 2 Analysis of risk classes. The chart shows patients' true risk classes as derived from the reference standard. For each risk class, the number of patients with identical risk classes (dark shading) and the number with different risk classes (light shading) are graphed. Solid triangles indicate the percentage of patients assigned to different classes. A different risk class was derived from the CPR data for 20 patients (13.3 percent). The proportion of misclassification was greatest in risk class 5.

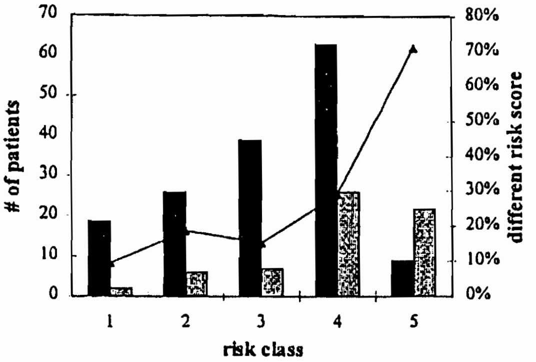


Figure 3 Analysis of risk scores by risk class. The chart shows patients' true risk classes as derived from the reference standard. For each risk class, the number of patients with identical risk scores (dark shading) and the number with different risk scores (light shading) are graphed. Solid triangles indicate the percentage of patients assigned to different classes. An identical risk score was derived from the CPR data for 163 patients (72.1 percent). In risk classes 4 and 5, the proportion of incorrect scores was greatest.

Results

In the 17-month study period, 226 of 241 inpatients met the inclusion criteria. Of the 15 excluded patients, 13 patients had a missing chief complaint, 1 patient was erroneously admitted to the emergency department without being seen by an emergency department physician, and 1 patient had both a missing chief complaint and was admitted erroneously to the emergency department instead of directly to the ward. The second chart audit revealed two discrepancies, when a coexisting disease was missed. The intrarater agreement for assessment of a patient's risk score was 92 percent.

The risk class computed for an individual patient from the CPR and from the Record was identical for 196 patients (86.7 percent). Among the remaining 30 patients (13.3 percent), the CPR-derived risk class was two risk classes lower for 1 patient (3 percent), one risk class lower for 22 patients (73 percent), and one risk class higher for 7 patients (20 percent). For one patient (3 percent), the CPR-derived risk class was the same because simultaneous errors (committed and omitted) equalized the two risk scores and the resulting risk classes. Twenty (87 percent) of the 23 patients with a lower CPR-derived risk class were assigned risk class 3, 4, or 5. Figure 2 summarizes the overall misclassification of the CPR-derived risk class stratification.

The risk score obtained from the CPR was identical for 163 patients (72.1 percent) and different for 63 patients (27.9 percent). Of the 63 patients with a different risk score, the score was lower for 50 patients (79 percent), equal for 1 patient (3 percent), and higher for 12 patients (19 percent). For the patient with an equal risk score, the points attributed to various errors resulted in an equivalent risk score. For patients who had a false risk score computed, a total of 78 errors occurred. A single error occurred for 52 patients, two errors for 8 patients, three errors for 2 patients, and four errors for 1 patient. Among 63 patients assigned risk class 3, 4, or 5, errors affected the risk score of 55 (87 percent) and accounted for 70 (90 percent) of all 78 errors. Figure 3 summarizes the overall scoring errors of the CPR stratified by risk class.

Among the 63 patients with a different risk score, a change of the risk class occurred for 29 (46 percent), of whom 6 patients had a higher risk class and 23 had a lower risk class derived from the CPR. For the remaining 34 patients, the deviant risk score did not influence the risk class. Of the 29 patients with a different CPR-derived risk class, 21 patients were in risk class 4 or 5.

The average availability was 0.991 for the 12 PSI data elements that were part of the standard data set and were required to be present. The number of missing values and the availability for each required variable

Table 3 ▪

Data Availability and Number of Missing Entries for Pneumonia Severity Index (PSI) Variables of the Standard Data Set for 226 Patients

Characteristic	Missing Values	Data Availability
Age	0	1.000
Gender	0	1.000
Current history	1	0.996
Past history	0	1.000
Medication	6	0.973
Heart rate	3	0.987
Respiratory rate	1	0.996
Systolic blood pressure	4	0.982
Temperature	3	0.987
Average		0.991

NOTE: The nine required variables of the standard data set cover 12 data elements of the PSI.

Table 4 ▪

Data Availability and Number of Missing Entries for Pneumonia Severity Index (PSI) Variables That Are Not Part of the Standard Data Set for 226 Patients

Characteristic	Missing Values	Data Availability
Oxygen saturation	23	0.898
Mental status	37	0.836
Blood urea nitrogen	14	0.938
Sodium	8	0.965
Glucose	14	0.938
Hematocrit	9	0.960
Arterial pH	84	0.628
Pleural effusion	195	0.137
Average		0.788

NOTE: Although the clinical parameters are needed for the evaluation of the PSI, the variables are not obtained for every pneumonia patient as part of clinical care in the emergency department.

are shown in Table 3. For the eight variables that were not part of the standard data set, the average availability was 0.788 for the 226 patients. The number of missing values and the data availability are shown in Table 4. Arterial pH and pleural effusion were the variables with the lowest data availability (0.63 and 0.14, respectively). Blood gas analysis was not performed for 84 patients, and the presence of a pleural effusion was noted for 31 patients. The data availability for the four optional PSI variables that were results of laboratory tests was 0.94 or higher. Missing data variables yielded an error in the risk scores of 17 patients (7.5 percent).

The concordance of data variables was assessed without the finding "pleural effusion." The average concordance for the remaining 19 PSI variables in the 226 patients was 0.982 (4,216 concordant characteristics divided by the total number of 4,294 characteristics). For the 19 considered variables, the average number of concordant variables per patient was 18.65.

Different risk scores originating from the free-text fields accounted for 50 errors (64 percent), of which 34 were attributed to the emergency department nurse charting and 16 to the parsing algorithm. The 34 free-text errors occurred because a PSI-relevant coexisting disease was not charted for 22 patients, and 12 patients were not identified as coming from a nursing home. The 16 errors caused by an imprecise parsing algorithm were due to misspellings for 3 patients, an incomplete and imperfect list of query terms for 12 patients, and the misspelling of a phrase that simultaneously represented a missing query term for one patient. Different risk scores caused by errors in the coded data element ("mental status," "gender") accounted for 16 errors (21 percent), all of which were ascribed to "mental status" and none to "gender." All the errors in coded data were attributed to the emergency department nurse. One pathologic laboratory finding (1 percent) was missed because the patient failed outpatient treatment for pneumonia and was admitted to the hospital the following day with an abnormal blood urea nitrogen value. Because the patient's type of encounter was converted from outpatient to inpatient registration, the abnormal blood urea nitrogen value was the second measurement for the patient's encounter. Among nine patients, there were 11 errors (14 percent) in the vital signs, of which

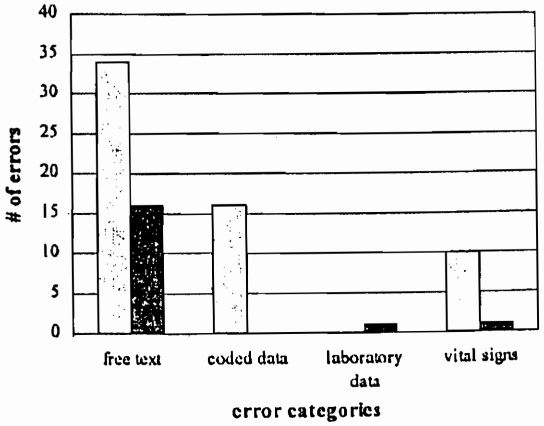


Figure 4 Analysis of types of errors and their causes. The most prevalent type of error occurred in the free-text category and accounted for 64 percent of all errors. The laboratory data had the most correct data. The cause of error was attributed to emergency department nurse charting (light shading) for 77 percent of errors and to system errors (dark shading) for the rest.

10 were attributable to the emergency department nurse and one was caused by a rounding error. For six patients the emergency department nurse did not chart the oxygen saturation under room air condition, and for two patients no vital signs were charted, which accounted for four missing values.

Errors attributable to emergency department nurse charting accounted for 77 percent of the total, and the remaining 23 percent were categorized as system errors. Figure 4 summarizes the type and cause of errors that resulted in a different PSI risk stratification.

Discussion

This study investigated the quality of data routinely available in the HELP System to drive the PSI guideline in real time and to deliver decision support at the point of care. Although the PSI was developed following rigorous methodologic requirements,[22,27] the complexity remains high, jeopardizing its dissemination and implementation. Despite its complexity, the PSI algorithm fulfills the criteria for a successful computerization, including the presence of clear definitions of variables and a decidable algorithm.[28] However, a CPR that accommodates the PSI guideline should meet additional criteria, such as the presence of sufficient and routinely available data and a high level of data quality.

We analyzed data quality from the clinical perspective on computerizing the severity index. The clinically relevant piece of information that is finally presented to the clinician is the patient's risk class. An identical risk class was obtained from the CPR for 86.7 percent of patients. This result is encouraging, considering that the paper-based version of the guideline is too complex to be easily memorized, many data variables originate from nurse charting, no additional data entry was required to achieve this level of data quality, and information from free-text fields was included. Presenting the CPR-available PSI information at the time of decision making enables clinicians to improve the accuracy of the risk class by correcting errors or adding missing information.

The availability of the PSI variables included in the standard data set is sufficiently high to run the respective part of the PSI from the CPR. In one third of the 18 incomplete records, the emergency department nurse left the entry for "current medication" empty. The emergency department nurse is supposed to enter "none" in a free-text field if the patient does not report any pertinent information. A charting practice that leaves free-text fields blank introduces ambiguity in the interpretation, because it is not known whether the patient is actually not taking any medication or whether the emergency department nurse forgot to chart the information or even ask the patient for it.

The interpretation of the data availability for the optional PSI variables is difficult because the data were evaluated retrospectively and were compared with information in the patient's record. If the emergency department physicians choose to apply the PSI guideline routinely, they would be required to obtain an oxygen saturation, laboratory data, and a blood gas analysis for every pneumonia patient who did not fall into the lowest risk class. In a prospective analysis of the PSI, we would predict that the data availability of the optional PSI variables would be even higher.

The analysis of the risk score presents a different perspective and is a direct result of the data quality. Although the average concordance of 0.98 for variables that are part of the standard data set appears to be high, the 78 errors considerably affected the risk scores. More than one quarter (27.9 percent) of the CPR-derived risk scores were different and underestimated the real risk for most patients. The underestimation of the CPR-derived risk score raises a problem for a computerized implementation. The calculated risk score might translate into a lower risk class, underestimating the patient's true risk of mortality. If the risk class is used as an admission criterion,[21] risk classes that are too low mean that more patients are treated as outpatients when they actually should be admitted to the hospital.

The most frequent errors occurred in the free-text category, and the most common cause of these errors was inaccurate nurse charting. Errors originating from the free-text fields might be reduced by applying a more sophisticated parsing algorithm, especially as natural language processing methods become available and are incorporated into CPRs. Free-text fields remain difficult to use in decision support systems. An alternative approach to increasing data quality is to try to encode the terms that appear most frequently in the free-text fields.

The interpretation of the risk scores, however, must be viewed in terms of the conservative approach of the study design. Of the 372 inpatients and outpatients who were diagnosed with pneumonia in the emergency department during the study period, we included only the 241 inpatients. In general, patients admitted to the hospital have more coexisting diseases and more abnormal findings than outpatients. To evaluate the PSI computerization, we focused on inpatients only, because outpatients present with few abnormal findings. The computerized records of inpatients provide more opportunities for the commission of errors in the computerization process than do the records of outpatients.

It is surprising that all 20 variables needed to compute the PSI during a patient's encounter in the emergency department are available in electronic form which promotes an automation of the PSI. However, it is important that all the variables of the PSI or of any other guideline are present in computable format and do not require additional data input. The emergency department is a busy clinical setting for the implementation of guidelines,[4] emphasizing the importance of complete automation of the PSI.

A limitation of this study is that one author performed both the CPR evaluation and the review of the patient's Record. Ideally, persons blinded to the purpose of the study would carry out both the CPR evaluation and the review of Records. We attempted to minimize observer bias by standardizing the data collection with a self-coding data sheet and by performing a second chart audit of 10 percent of randomly selected patient Records.

Another limitation concerns the vital signs. For clinical ease, the PSI dichotomizes the continuous values such as heart rate or oxygen saturation into normal and abnormal values. Therefore, we determined whether the actual value was identical not on a continuous scale but on a dichotomous scale. For example, a heart rate value was considered normal even if a data entry error occurred when the nurse incorrectly charted the rate at 60 beats/min instead of 80 beats/min. The majority of data entry errors are impossible to detect because, in contrast to the nurse charting notes, the dictated reports do not contain a time stamp indicating when values were recorded. To resolve discrepancies between a normal value obtained from the CPR and an abnormal value obtained from the Record, the abnormal value was considered the correct one. We did not quantify data entry errors that may have occurred when values from both the CPR and the Record were normal. However, data entry errors seem not to be an important cause of incorrect data.[19]

To ease and promote its clinical application, the PSI was simplified from a logistic regression to a scoring algorithm. Computerizing the PSI on the basis of the original logistic regression takes advantage of the computational power of a clinical information system and provides a probability and a 95 percent confidence interval. For clinicians, the probability may represent a more precise and intuitive mortality measure than the less meaningful PSI risk class.

The computerization of paper-based guidelines is desirable, because it assists health care providers with easily accessible decision support at the point and the time of care. The computerized representation of the validated and clinically useful PSI guideline supports the implementation and dissemination of the prediction rule. Although the paper-based PSI guideline has advantageous characteristics for an automation, successful clinical implementation depends on the availability of high-quality data. The level of data quality should be assessed prior to implementation, because identification of the sources of errors supports efforts to improve data capture and provide correct and complete data. Enabling clinicians to review and modify the data variables used to generate guideline suggestions represents a possible approach to achieving an accurate risk measure. This approach comes at the cost of additional data entry, however, and it remains uncertain whether clinicians are willing to trade additional data entry for a higher level of data quality in the CPR. Demonstrating a high level of data quality prior to guideline implementation increases the credibility toward computer-generated guideline recommendations and ensures that clinicians can eliminate existing inaccuracies in the risk assessment with few corrections or additions. Only implementation of the PSI as an integrated computerized decision support tool will indicate whether the automated recommendations will influence the clinician's decision making.

The authors thank Wendy Webber Chapman, Nathan C. Dean, MD, and Charles Lagor, MD, for helpful comments in the preparation of the manuscript.

References ■

1. Elson RB, Connelly DP. Computerized patient records in primary care: their role in mediating guideline-driven physician behavior change. Arch Fam Med. 1995;4:698–705.
2. Zielstorff RD. Online practice guidelines: issues, obstacles, and future prospects. J Am Med Inform Assoc. 1998;5:227–36.
3. Schriger DL. Emergency medicine clinical guidelines: we can make them, but will we use them? Ann Emerg Med. 1996;27:655–7.
4. Berger JT, Rosner F. The ethics of practice guidelines. Arch Intern Med. 1996;156:2051–6.
5. Barnett GO, Winickoff RN. Quality assurance and computer-based patient records. Am J Public Health. 1990;80:527–8.
6. Nilasena DS, Lincoln MJ, Turner CW, et al. Development and implementation of a computer-generated reminder system for diabetes preventive care. Proc Annu Symp Comput Appl Med Care. 1994:831–5.
7. Elson RB, Connelly DT. Applications of computer-based clinical guidelines. JAMA. 1998;279:989–90.
8. Lobach DF, Hammond WE. Computerized decision support based on a clinical practice guideline improves compliance with care standards. Am J Med. 1997;102:89–98.
9. Tierney WM, Overhage JM, Takesue BY, et al. Computerizing guidelines to improve care and patient outcomes: the

example of heart failure. J Am Med Inform Assoc. 1995;2: 316–22.

10. Evans RS, Pestotnik SL, Classen DC, et al. A computer-assisted management program for antibiotics and other anti-infective agents. N Engl J Med. 1998;338:232–8.
11. Morris AH, Wallace CJ, Menlove RL, et al. Randomized clinical trial of pressure-controlled inverse ratio ventilation and extracorporeal CO_2 removal for adult respiratory distress syndrome. Am J Respir Crit Care Med. 1994;149:295–305.
12. Tierney WM, Overhage JM, McDonald CJ. Toward electronic medical records that improve care. Ann Intern Med. 1995;122:725–6.
13. Zielstorff RD, Barnett GO, Fitzmaurice JB, et al. A decision support system for prevention and treatment of pressure ulcers based on AHCPR guidelines. Proc AMIA Annu Fall Symp. 1996:562–6.
14. Schriger DL, Baraff LJ, Rogers WH, Cretin S. Implementation of clinical guidelines using a computer charting system: effect on the initial care of health care workers exposed to body fluids. JAMA. 1997;278:1585–90.
15. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA. 1998;280:1339–46.
16. Tierney WM, Overhage JM, McDonald CJ. Computerizing guidelines: factors for success. Proc AMIA Annu Fall Symp. 1996:459–62.
17. Classen DC. Clinical decision support systems to improve clinical practice and quality of care. JAMA. 1998;280:1360–1.
18. Jacobson PD. Legal and policy considerations in using clinical practice guidelines. Am J Cardiol. 1997;80(8B):74H–79H.
19. Wagner MM, Hogan WR. The accuracy of medication data in an outpatient electronic medical record. J Am Med Inform Assoc. 1996;3:234–44.
20. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc. 1997;4:342–55.
21. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med. 1997;336:243–50.
22. Wasson JH, Sox HC. Clinical prediction rules: have they come of age? JAMA. 1996;275:641–2.
23. Kuperman GJ, Gardner RM, Pryor TA. HELP: A Dynamic Hospital Information System. New York: Springer-Verlag, 1991.
24. Gardner RM, Pryor TA, Warner HR. The HELP hospital information system: update 1998. Int J Med Inf. 1999;54:169–82.
25. Haug PJ, Gardner RM, Evans RS. Hospital-based decision support. In: Berner ES (ed). Clinical Decision Support Systems: Theory and Practice. New York: Springer-Verlag, 1998.
26. Fine MJ, Hanusa BH, Lave JR, et al. Comparison of a disease-specific and a generic severity of illness measure for patients with community-acquired pneumonia. J Gen Intern Med. 1995;10:359–68.
27. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: applications and methodological standards. N Engl J Med. 1985;313:793–9.
28. McDonald CJ, Overhage JM. Guidelines you can follow and can trust: an ideal and an example. JAMA. 1994;271:872–3.

CHAPTER 5

EVALUATION OF A COMPUTERIZED DIAGNOSTIC DECISION SUPPORT SYSTEM FOR PATIENTS WITH PNEUMONIA: STUDY DESIGN CONSIDERATIONS

Dominik Aronsky, MD, Peter J Haug, MD

J Am Med Inform Assoc. 2000 (submitted)

5.1 Abstract

Planning the clinical evaluation of a computerized decision support system requires an optimal strategy that unifies the different aspects of the clinical problem, the technical difficulties of software and hardware integration and implementation, the behavioral aspects of the targeted users, and the discipline of study design. Although clinical information systems are becoming more widely available, only a few decision support systems have been formally evaluated in a clinical environment. Thus, detailed experiences about the difficulties in the clinical evaluation of a decision support system remain scarce. We report a variety of important barriers that we were required to address while planning a clinical trial for the evaluation of an integrated, real time decision support system for the automatic identification of patients likely to have pneumonia in an emergency department. The challenges involved behavioral, logistical, economical, technical, clinical, and work flow issues, and influenced our decision making process in choosing an optimal study design. In the absence of a true gold standard, we illustrate how we created a credible and clinically acceptable reference standard for pneumonia with limited financial resources. For the creation of a reference standard, we describe the importance of recognizing verification bias and how introducing verification bias can be avoided. Finally, advantages and disadvantages of different study designs are explored with respect to the targeted users and the clinical setting.

5.2 Introduction

The development and evaluation of diagnostic decision support systems (DDSS) remains an active and challenging area of research. Several DDSS have demonstrated promising diagnostic performances in formal evaluations (1-4). However, the majority of diagnostic systems have not been evaluated in a clinical environment. Others have been evaluated on a limited set of patients only. The few prospectively evaluated clinical systems were stand-alone systems (3, 4) and were not integrated into clinical information systems.

Stand-alone systems depend heavily on users to enter data. They generally include data elements that are not routinely documented in the clinical information system during a patient's encounter or are not captured in an easy computable format. The limitations caused by additional, sometimes redundant data entry and time constraints during a patient encounter prevent health care providers from applying the diagnostic information for routine patient care.

Integrating a DDSS into an existing clinical information system and into the workflow of busy clinicians is desirable. Separate and redundant data entry can be reduced or eliminated, and the delivery of DDSS information can be blended more easily into the clinician's workflow. Integration provides an opportunity to apply the diagnostic information for a variety of computerized clinical applications. For example, the diagnostic information might be applied to trigger disease specific guidelines automatically and independent of a physicians' initiation. Without an automatic detection process, identifying patients who are eligible for a disease specific guideline, either paper-based or computerized, remains the responsibility of the health care provider. The implementation of computerized guidelines remains a challenge, and integrated decision support systems might support the implementation efforts. Clinicians may consider automatically displayed guideline information more often when the information is readily available information with little effort. Evaluating a paper-based guideline requires clinicians to spend additional time to complete the guideline form.

The challenge of diagnostic accuracy exists for both stand-alone and integrated DDSSs. However, the clinical evaluation of an integrated DDSS creates new challenges that do not exist or have less an influence in the evaluation of a stand-alone DDSS. Examples for challenges not present in the evaluation of stand alone systems include the data quality of a clinical information system, the format and timely availability of patient data, and the location and timing of delivering the DDSS information. Recognizing the possible influential factors supports the decision making process in choosing an appropriate strategy for a clinical evaluation. Experiences in designing a study that prospectively evaluates an integrated real-time DDSS in a clinical setting are scarce.

In this paper we describe the design for the prospective clinical evaluation of a real time, integrated DDSS for patients with community-acquired pneumonia. First, we describe the clinical setting and the functional characteristics of the population based DDSS. We address the difficulties in establishing a clinically acceptable reference standard to determine the system's overall accuracy and highlight the importance of recognizing the presence of verification bias. We illustrate the tradeoffs when choosing a clinical characteristic, such as a patient's chief complaint, as a preselection mechanism for increasing the number of patients having the target disease. We describe the advantages and disadvantages of different study design alternatives and discuss why certain decisions in the study design were considered preferable to other design options. We discuss how the expected behavior of the users, the clinical workflow, and the planned intervention would have influenced different study designs.

5.3 A Computerized Diagnostic Decision Support System for Pneumonia

We developed and implemented a pneumonia DDSS for use in the emergency department (ED) of LDS Hospital, a 520-bed university-affiliated tertiary care center in Salt Lake City, Utah (5, 6). Our main objective was to develop a real-time process that automatically identifies ED patients who present with findings suggestive of pneumonia. The system then triggers the computerized evaluation of a patient and disease specific pneumonia guideline. For this purpose the system consists of a diagnostic and a disease management component. The diagnostic component is based on a probabilistic algorithm (Bayesian network) that computes a probability of pneumonia. The disease management component consists of the Pneumonia Severity of Illness Index (PSI) which is computed for patients likely to have pneumonia. The PSI calculates a risk score based on twenty routinely available, computer-charted variables in the ED and stratifies patients into five risk classes (7). The PSI risk classes can be applied to support clinicians in the admission decision (7, 8).

To eliminate additional data entry and allow a high level of integration into the clinical information system the DDSS was developed with data elements that were routinely collected during the patient's ED encounter. In addition, almost all data elements required for the PSI were routinely captured and stored in our clinical information system. Prior to system implementation the accuracy of the PSI was assessed when data elements from the clinical information system were used and resulted in accurate risk class for 86 percent of admitted pneumonia patients (9). Taking advantage of routinely available data elements in the clinical information system allows the decision support system to update and display the probability for pneumonia and the PSI risk class information in an unsolicited manner. Rather than leaving the responsibility of initiating the PSI calculation in the hands of busy clinicians, the DDSS identifies patients likely to have pneumonia and supports the automatic delivery of the PSI information at the point and time of care.

The ED main screen displays a list of current ED patients and represents the most common entry point for charting and accessing patient information. In addition to basic patient information, such as patient names, vital signs, and the availability of laboratory values or dictated reports, a dedicated column provides space for the results of protocols (Figure 5.1). Displaying the pneumonia related information on the top level ED screen assures that the information is available and can be easily located and seen by clinicians.

The clinical information system provides the data elements and is used to display the pneumonia information. However, the ED staff members are not able to examine more detailed information about the pneumonia probability or the PSI. More detailed information is available on a separate, dedicated computer in the ED. The computer gives ED physicians who are interested in the variables involved in calculating the pneumonia probability or the PSI, the opportunity to review, add, correct, or delete patient information. If the physicians make changes, the system immediately updates the probability and the risk mortality model reflecting the changes in the patient's findings.

5.4 The Influence of Disease Prevalence

The disease prevalence at the developing site directly influences the diagnostic characteristics of a decision aid (10, 11). The disease prevalence appears to influence the positive and negative predictive values more than the sensitivity and the specificity (12). For clinical purposes the predictive values are more useful because they inform the clinicians about the expected proportion of diseased patients when the test is positive (or negative). Although tests that are applied in low prevalence diseases can yield relatively high sensitivity and high specificity, the positive predictive value usually remains at a moderate level. The moderate positive predictive value results from the heavily unbalanced distribution of the cell frequencies in a 2x2 contingency table. With decreasing disease prevalence the distribution in a 2x2 table becomes even more unbalanced. Table 5.1 illustrates how disease prevalence influences the predictive values if sensitivity and specificity are kept constant.

The prevalence of pneumonia in our ED population averages about 1.7 percent and fluctuates due to seasonal variations. During the winter months pneumonia affects patients more frequently, and the disease prevalence may substantially increase (2.7 percent in our setting). During the summer months pneumonia is less frequently seen and the prevalence may be low (0.8 percent in our setting). It was our goal to develop a DDSS that operated similarly to a screening test and would be able to identify pneumonia patients from an unrestricted population, such as the entire ED population. Because we choose to develop a system that could be applied to an entire population, the disease prevalence was small and consequently a moderately low predictive value for the DDSS was found during the development phase. The low predictive value of the DDSS had implications for displaying PSI guideline information. PSI information was expected to be displayed in many patients without pneumonia, but who had a disease with similar clinical presentation, such as acute bronchitis, congestive heart failure, or pulmonary

Table 5.1: Influence of Disease Prevalence on Test Characteristics with Constant Sensitivity (90%) and Specificity (75%).

5.1a Disease Prevalence = 20 %

	disease		
	present	absent	total
test pos	180	200	380
test neg	20	600	620
total	200	800	1000

pos predictive value	47.4%
neg predictive value	96.8%

5.1b Disease Prevalence = 10 %

	disease		
	present	absent	total
test pos	90	225	315
test neg	10	675	685
total	100	900	1000

pos predictive value	28.6%
neg predictive value	98.5%

5.1c Disease Prevalence = 2 %

	disease		
	present	absent	total
test pos	18	245	263
test neg	2	735	737
total	20	980	1000

pos predictive value	6.8%
neg predictive value	99.7%

embolism. We discussed the low positive predictive value with the ED physicians. For clinical purposes they considered the low positive predictive value acceptable, specifically as the DDSS information was displayed automatically and did not require additional data entry. The ED physicians preferred an automatic approach with lower predictive power to an approach that demonstrated higher predictive power but required data entry.

5.5 The Influence of Preselection Criteria

A possible method to increase the predictive value and reduce unnecessary information consists of using a prescreening factor that reduces the number of patients from the underlying population. Depending on the test characteristics, however, the limitation of eligible patients based on the prescreening factor might show no or even an opposite effect. If an increase in the positive predictive value is achieved, introducing a selection criterion may come at the cost of selection bias, and the obtained results are limited to the prescreened subpopulation. Additionally, preselecting patients based on certain clinical criteria may limit the population to patients with typical findings of the disease. However, clinicians rarely need decision support for patients presenting with typical findings, but for patients with uncommon findings.

Typical findings for pneumonia patients include a respiratory complaint, fever, and cough. However, pneumonia is a frequent disease in the elderly population, and elderly patients often do not present with the typical findings (13). Elderly patients may not have a fever or a cough, but have a change in mental status, a syncopal episode, or general malaise. Other patients may complain about abdominal pain or a headache. The diagnostic challenges occur in the patients with atypical findings, and clinicians might be expected to benefit from a DDSS most in these patients.

To examine these effects we explored whether a preselection procedure used on the ED population based on the patient's chief complaint influences the positive predictive value of our system. The most frequent chief complaint in our historical data set was a respiratory symptom and accounted for 55 percent of all pneumonia patients. However, pneumonia patients may complain about numerous other chief complaints (Table 5.2). Eleven other chief complaints accounted for an additional 40 percent of pneumonia patients. The remaining 5 percent included an additional 14 chief complaints.

Although a DDSS may perform extremely well in patients with the typical respiratory chief complaint, clinicians are usually not challenged and do not require computer support in many patients with typical findings. Preselecting patients based on the chief complaints did not increase the positive predictive value but had an opposite effect on the specificity and the negative predictive value of the DDSS (Table 5.3). We believed that the DDSS would change physicians' behavior more often in patients with uncommon and less frequent chief complaints such as abdominal pain, headache, or weakness. For patients presenting with the typical pneumonia findings, we assume that the busy ED physicians will focus on the PSI risk class rather than the pneumonia probability.

5.6 Reference Standard and Verification Bias

We developed the DDSS with historical data from more than 32,000 ED patients at LDS Hospital. For the DDSS development we identified pneumonia patients by ICD-9

Table 5.2: Chief Complaints for Pneumonia Patients

Chief complaint	Pneumonia patients		all ED patients	
	Abs %	cum %	abs %	cum %
Respiratory	54.7%	54.7%	10.4%	10.4%
Fever	14.5%	69.2%	3.4%	13.8%
Chest pain	10.9%	80.1%	14.1%	27.9%
Abdominal	3.0%	83.1%	8.1%	36.0%
Neurological	2.4%	85.5%	4.9%	40.9%
Abdominal pain	2.2%	87.7%	18.1%	58.9%
Falls	1.6%	89.3%	8.6%	67.6%
Weak(ness)	1.6%	90.9%	0.9%	68.5%
Body Aches	1.4%	92.4%	0.8%	69.3%
Temperature Related	1.0%	93.4%	0.2%	69.5%
Cardiovascular	1.0%	94.4%	3.1%	72.6%
Ear/Nose/Throat	1.0%	95.4%	3.2%	75.8%

Table 5.3: Using the Chief Complaint as a Preselection Criterion: Influence on Test Characteristics With a Constant Sensitivity of 95 Percent

	pneumonia patients included	pneumonia prevalence	positive predictive value	negative predictive value	specificity
most frequent chief complaint	55%	10.4%	15.5%	98.8%	44.4%
three most frequent chief complaints	80%	6.2%	14.8%	99.6%	67.6%
eight most frequent chief complaints	91%	2.9%	13.1%	99.8%	83.1%
all chief complaints	100%	1.7%	13.1%	99.9%	90.7%

discharge diagnosis. We recognized that diagnostic information obtained from claims data are known to be imperfect and imprecise for clinical purposes (14, 15). However, for the development of a population-based DDSS from a large clinical database, ICD-9 codes were the most feasible and economical diagnostic information. For a clinical evaluation, however, the use of claims data is inferior.

For the diagnosis of pneumonia, no objective definition, such as for the diagnosis of an acute myocardial infarction (4) exists. The most frequently applied definition for pneumonia is the presence of a clinical finding suggestive of pneumonia and the identification of a new infiltrate on the patient's chest x-ray during the patient's initial presentation (7). However, observer variation in physical examination and radiological interpretation occur (16, 17). A positive microbiology culture represents a strong indicator for the presence of pneumonia. However, a negative microbiology result does not indicate the absence of the disease, and the microbiologic cause remains unknown in 94.3 percent of outpatients and in 71.4 percent of inpatients (18). Even if more rigorous criteria, such as histologic lung tissue, are available, pathologists may disagree about the presence of pneumonia (19). In addition, several clinical situations may occur that confirm a pneumonia diagnosis, but do not meet the above criteria. For example, a pneumonic infiltrate may appear not on the day of presentation, but during the following day. Another example is that a pneumonic infiltrate in the lower lung lobes is diagnosed on an abdominal film or a chest computer tomogram. Clinicians may choose not to perform a chest x-ray for patients who have adequate radiological evidence for an infiltrate on a radiologic exam other than a chest x-ray.

Consequently, for the prospective clinical evaluation of the DDSS we were faced with two important problems. The first problem is well recognized and concerns the creation of a solid and credible reference standard for pneumonia. The second problem is less frequently recognized and addresses verification bias.

The presence of a gold standard diagnosis forms the backbone for evaluating a DDSS. Specifically for a patient's diagnosis, the best available gold standard is definitely preferred, but may be difficult and expensive to obtain. In medical informatics adopting a "silver" reference standard that is feasible, economical, and ethical is sometimes necessary (20). If the definition of a "silver" reference standard becomes necessary for the evaluation of a decision support system in a clinical setting, the standard should at least be clinically acceptable. However, for other diseases, such as pneumonia, the lack of objective criteria to establish the diagnosis is approximated by more subjective measures. Thus the creation of a reference standard for pneumonia includes subjective physicians' judgments.

While planning the creation of an optimal reference standard, evaluators have to be aware that the work-up for confirming the patient's diagnosis does not introduce verification bias. Verification bias occurs if a test or a criterion is applied to patients whose disease status will subsequently be confirmed by a gold standard procedure. Selecting patients for verification of disease status with respect to a positive test result only introduces bias. Patients with a negative test do not undergo the gold standard procedure and the true disease status remains unknown (21, 22). The disease status of all patients in the study population needs to be confirmed by the gold standard independent on the result of the test, because the final diagnosis is determined by the gold standard procedure and remains unknown until the gold standard procedure is applied to all study patients.

An example for verification bias is the work-up of patients suspected to have a pulmonary embolism. To assess the test characteristics of ventilatory-perfusion scans for identifying patients with a pulmonary embolism, pulmonary angiography was commonly used as the gold standard procedure to verify the disease status. Pulmonary angiography bears more risks for the patient than ventilatory-perfusion scans. If only patients with a positive ventilatory-perfusion scan undergo pulmonary angiography, verification bias is present, because not all patients have the chance of being submitted to the gold standard procedure, i.e., the disease status of patients with a negative ventilatory-perfusion scan are not verified and assumed to be negative. It is possible that a negative ventilatory-perfusion scan is a false negative result and pulmonary embolism is present, but the scan failed to detect the disease. These patients are classified as not having pulmonary embolism, because their disease status is not verified by the pulmonary angiography, the chosen gold standard for pulmonary embolism. To avoid verification bias, all patients with suspected pulmonary embolism need to be submitted to both the test (ventilatory perfusion scan) and the gold standard (pulmonary angiography). In such a design, the test characteristics of ventilatory-perfusion scans can be determined exactly and depend only on the quality of the gold standard to identify the true disease status of patients with pulmonary embolism.

Verification bias may considerably influence the outcome measures by either inflating the sensitivity and deflating the specificity, or by showing an effect in the opposite direction. Verification bias can be avoided when patients are selected for the gold standard procedure based on the disease and not on the outcome of the test. In studies with verification bias a correction procedure can be applied if underlying information about the selection procedure is known (21). In diseases with assured clinical manifestation (e.g., appendicitis), verification bias can be avoided by following up patients assumed negative.

For the evaluation of the pneumonia DDSS, verification bias was initially present in a subtle way. In the historical data set, pneumonia patients were identified by applying ICD-9 codes. Accepting ICD-9 codes as a reference standard for the diagnosis of pneumonia has the advantage that every ED patient is assigned a diagnostic code. Having a code for all patients divides the population into patients with and without pneumonia. Because every patient obtains the same work-up (in the form of ICD-9 codes) verification bias is not present. However, as mentioned above, a well-documented disadvantage of ICD codes is the inaccuracy in the coding procedure. To improve the coding deficiencies, cases with an ICD-9 code of pneumonia could be submitted to physicians for an in depth review. This approach, however, quietly introduces verification bias because not all ED patients have an equal chance to obtain the reference standard verification in the form of an in depth review. Specifically pneumonia patients with ICD-9 codes different from pneumonia would be completely missed. Verification bias is present because we would select patients based on the test (ICD-9 code) and not based on the patient's disease (pneumonia). The DDSS evaluation would yield an inaccurate sensitivity and specificity. In the historical data set, we used ICD-9 codes for identifying pneumonia patients. Consequently, verification bias existed and the obtained results have to be interpreted with caution.

Unconditionally reviewing all ED patients during the study period avoids verification bias. However, 60 to 70 patients are treated in our ED daily and a detailed review of each patient's chart is uneconomical and unfeasible due to our limited financial and personal resources. Before performing the prospective study we sought a method that both controls effectively for verification bias and remains feasible to perform within the limits of our resources. We defined a three-step process that verified the disease status of ED patients. The first step applied five different criteria: (1) the patient's chief complaint, (2) the presence of a radiology chest examination, (3) the patient's ICD admission and discharge codes, (4) a keyword search, and (5) the DDSS's computed pneumonia probability. Although we applied criteria that potentially introduce verification bias, the goal was not to identify patients with pneumonia, but to safely exclude patients who have only a very remote chance of pneumonia. It is highly improbable that a patient with pneumonia did not meet any one of the five criteria.

5.7 Creating a Reference Standard for Pneumonia

The chief complaint is part of the triage assessment, and the ED nurses enter the chief complaint in a coded format in >98 percent of the patients. Based on the patients' chief complaints from our development data set we determined all the chief complaints about which at least one pneumonia patient had complained of. All ED patients with these chief complaints were included for second step. We included all ED patients who had a chest radiology exam performed, even if the patient's chief complaint did not occur in the historical data set. Furthermore we included all patients who had an admit or discharge ICD-9 diagnosis of pneumonia. Patients with a probability of more than one percent, as computed by the DDSS, were also included. Finally we performed a key word search in the ED physicians' report, the admission report, and the discharge report, and included the patients who had the term "pneumonia" mentioned in any of the reports, specifically in any of the ED follow-up reports. Based on an analysis of our historical data set we estimated that up to 73 percent of the ED patients could be excluded based on the first step. The actual exclusion rate was slightly lower (66.7 percent).

In the second step, a group of five physicians read the ED physicians' reports and the radiologists' reports of the chest exams. The group included a second-year resident, two third-year residents, and two board certified internists. The reviewers did not know what criteria were applied to patients to be reviewed. The reviewer's task was to exclude patients who had no chance of pneumonia from further classification efforts. However, the physicians were instructed not to make a decision whether pneumonia was present. We estimated that the second step would reduce the number of patients to be reviewed in the third step by an additional 21.6 percent. The actual elimination rate in step two was 25.6 percent.

Patients who were considered to have any small remaining chance of pneumonia in the second step were thoroughly reviewed in the third step. The third step involved a review of the patient's chart and radiology exams by physicians who were board-certified in pulmonary and critical care medicine. At least two different physicians reviewed each patient's information. If the two reviewers disagreed, a third physician reviewed the case. The majority vote decided whether pneumonia was present or absent. The reviewers not only differentiate whether pneumonia was present or absent, but categorized the type of pneumonia according to established criteria (23). The types of pneumonia included community-acquired pneumonia, hospital-acquired pneumonia, pneumonia in an immuno-compromised patient, suspected aspiration pneumonia, pneumonia due to tuberculosis, and postobstructive pneumonia due to malignancy. We estimated that about 5 percent of all study patients remained to be reviewed in the last step. Five percent represented about three times the number of expected pneumonia patients during the study period. The physicians reviewed 7.7 percent of the entire study population. A summary of estimated and actual rates is shown in Table 5.4. In the last step two reviewers agreed on the presence or absence of pneumonia in 89 percent (85 percent estimated) of patients. The disagreement was resolved by the third reviewer in 11 percent (15 percent estimated) of cases. This approach is an economical method to create a majority vote without having all three reviewers judge all patients.

In addition to assessing the absence or presence of pneumonia, step three reviewers were asked to judge whether the diagnosis made at the patient's initial ED encounter was "correct," "suspected," "missed," or "incorrect." The category "correct" was marked for patients whose ED diagnosis equaled the gold standard diagnosis. The category "suspected" was marked for the frequent situations in which a preliminary or working diagnosis, such as sepsis, fever of unknown origin, acute exacerbation of bronchitis, or a change of mental status in an elderly patient, was established. Pneumonia was not the ED physician's final assessment, but it was considered in the differential diagnosis list and was expected to be assessed during further patient work-up. "Missed" or false negative pneumonias were defined for gold standard pneumonia cases that did not include pneumonia in the final ED diagnosis or among the ED physicians' differential

Table 5.4: Comparison Between Estimated and Actual Accrual Rates

	estimates		actual	
	absolute	percent	absolute	percent
study patients	9300		10863	
pneumonia patients	155 [(1)]	1.7%	273 [(2)]	2.5%
patients / day	63		70	
study period	147		155	
physician review for reference standard: step two	2500	26.9%	3618	33.3%
physician review for reference standard: step three	490	5.3%	838	7.7%

[(1)] pneumonia diagnosis based on ICD-9 codes

[(2)] pneumonia diagnosis based on clinical review diagnoses. "Incorrect" or false positive pneumonias included patients with an ED diagnosis of pneumonia whose gold standard diagnosis was not considered to be pneumonia.

In evaluation studies that apply a reference standard it is advisable to assess reliability (interrater agreement) and repeatability (intrarater agreement) (20, 24). The reliability of a reference standard increases as more physicians participate in the review process. Due to economical considerations and the involved reviewers' limited time the third step in the creation of our reference standard involved only two physicians. Because two physicians reviewed each patient and the third physician was involved for resolving disagreement, reliability may suffer. However, this approach is more economical and maintains the feasibility of establishing a majority vote for the disease of each patient.

The physicians involved in establishing the reference standard were independent reviewers who were practicing medicine. No ED physician and no member of the development team were reviewers. Reviewers involved in step three were not involved in step two of the review process, and vice versa. All reviewers, except one, were from the hospital where the study was performed. Optimally, reviewers were blinded to the purpose of the study. We chose to inform the reviewers about the general purpose of the study to motivate them for the tedious and time-consuming task of reviewing charts. However, they were unaware of any details involving the development and the operational characteristics of the DDSS. Although it is desirable to completely separate the tasks of the DDSS developers, the users, and the reviewers, it is frequently not practical. The financial resource allocated for the evaluation was US$ 25,000, of which

US$ 18,000 was spent for the review process alone. It is possible that the proposed three-step process will not identify the true disease status for all patients. However, we considered the chances of missing a pneumonia patient to be small, and the three-step process represents a balanced tradeoff between clinically acceptable disease verification, optimal resource allocation, and feasibility.

5.8 Considerations for Selecting a Study Design

The objective of the clinical evaluation was to evaluate the system's diagnostic performance and to test whether automatically providing physicians with computerized pneumonia information represented a feasible and successful approach to deliver guideline information. The outcome measure for assessing the system's overall diagnostic performance was the area under the receiver operating characteristic curve (25). Although assessing the overall diagnostic performance in a prospective study is desirable, the evaluation does not yield information whether the diagnostic data are valuable for clinical purposes. The value of a DDSS can be assessed by comparing the diagnostic performance of physicians with and without the system's information.

Assuming that delivering pneumonia guideline information is only valuable if the information is available for the majority of pneumonia patients, we choose to concentrate on differences in diagnostic sensitivity between physicians with and without the system. Because ICD-9 codes were the only available diagnostic source and are inaccurate for clinical purposes, we realized that our estimates for planning the study were vague. We estimated that physicians using the system would identify about 10-15 percent more pneumonia patients than without the system. The rough estimate was the basis for computing sample size, power, and the duration of the study.

It is important to recognize that the system represents an approach to automatically provide physicians with PSI information. Even if the system identified the same patients as the physicians only, it might still be valuable because the physicians had access to guideline information, which they did not have previously.

A variety of designs have been applied to evaluation studies in medical informatics; however, the reasons involved in preferring one study design to another have been rarely discussed. The interdisciplinary characteristics of decision aids and the variations in the clinical environment create specific challenges and unique barriers that influence the design of a clinical evaluation. For our study we considered a quasi-experimental design, an independent group comparison, and a mixed design. Here we discuss the strengths and weaknesses of each design emphasizing the behavioral, technical, economical, and statistical facets.

5.8.1 Time-Series Design

A quasi-experimental time-series design would compare the diagnostic accuracy for pneumonia during two successive time periods. During the first period the ED physicians would not have access to the DDSS information, and the ED physicians' diagnostic characteristics would be compared against the reference standard. During the second period the DSS information would be available, and the diagnostic characteristics would again be compared with the reference standard. Although a quasi-experimental design is powerful in detecting existing differences, there are potential threats to internal validity such as historical events or maturation (26).

5.8.2 Multiple-Baseline Design

A multiple baseline design, e.g., intervention off-on-off-on, is a possible alternative, but at least doubles the study period. The "off" period between the two "on" periods should be long enough that the outcome measures approach the initial baseline. Differences in outcome measures between the "on" and the "off" periods shrink as the washout period is shortened. Smaller differences, however, make it more difficult to detect an effect even if it is actually present. In statistical terms, the probability of committing a type two error increases.

Sometimes the introduction of a DDSS results in a learning effect that represents an alternative explanation and weakens the outcome. Even when concluding that the DSS had a measurable impact, it remains unknown whether the DDSS itself or the focus on the disease, such as increased awareness, better documentation, or a possible Hawthorne effect, influenced the observed change. In summary, the control for possible confounding factors in time series designs is difficult and the causal relationship between the introduction of a DDSS and the observed change may remain.

5.8.3 Parallel-Group Design

In the traditional experimental or parallel-group design, ED physicians are randomly assigned to either the intervention or the comparison group. The random assignment of ED physicians to two different groups has the advantage of being less vulnerable to threats of internal validity. However, the consequences of behavioral, logistical, and technical issues need to be addressed because they raise concerns about the ability to restrict the DDSS information only to the physicians in the intervention group. Without being exhaustive we describe five possible problematic areas:

- Sharing log-ins among clinicians due to inconsistent log-out practices: The ED nurses and physicians access the information system for charting and reviewing patient data. Specifically the ED nurses spend considerable time interacting with the computer. Terminals are installed in every patient room and in the central working area of the ED. Up to 80 percent of the data processing tasks are performed in the central area even though the computer availability is limited to six terminals. The ED staff does not consistently log out after using the computers although logging out is easy and accomplished with a single keystroke. If the ED staff does not log out, sessions stay alive until a time-out occurs or another user starts working on the terminal. In the latter case the following staff person may use the clinical information system under a wrong user identifier. Shortening the time-out period may require the ED staff to tolerate multiple log-in operations to finish a single task.
- Sharing experiences about the DDSS: Physicians share their experiences and discuss patients on a daily basis. ED physicians who have a particular good or bad experience with the DDSS may influence the attitude of the their colleagues.
- Changing responsibility for patients: Due to shift changes more than one physician may take responsibility for a patient. If, however, the two physicians were assigned to different groups of the experiment, it remains unclear whether the patient was diagnosed by a physician of the intervention or the comparison group.

- Attitude towards computers: Factors that are independent of the information presented by the DDSS, such as different preferences in or attitudes toward using computers in general might impact the evaluation. With only half of the ED physicians in the intervention group, the influence of individual attitudes grows. In our situation, the 12 ED physicians had a comparable amount of experience with the clinical information system.
- Physicians' preferences: It is conceivable that physicians have a professional interest in distinct diseases or conditions such as surgical patients, patients with respiratory symptoms, or elderly patients. Although it is unknown whether the effect is present in our ED physician group, it may influence the number of pneumonia patients in one of the two groups and represent a bias that is not controlled by randomization.

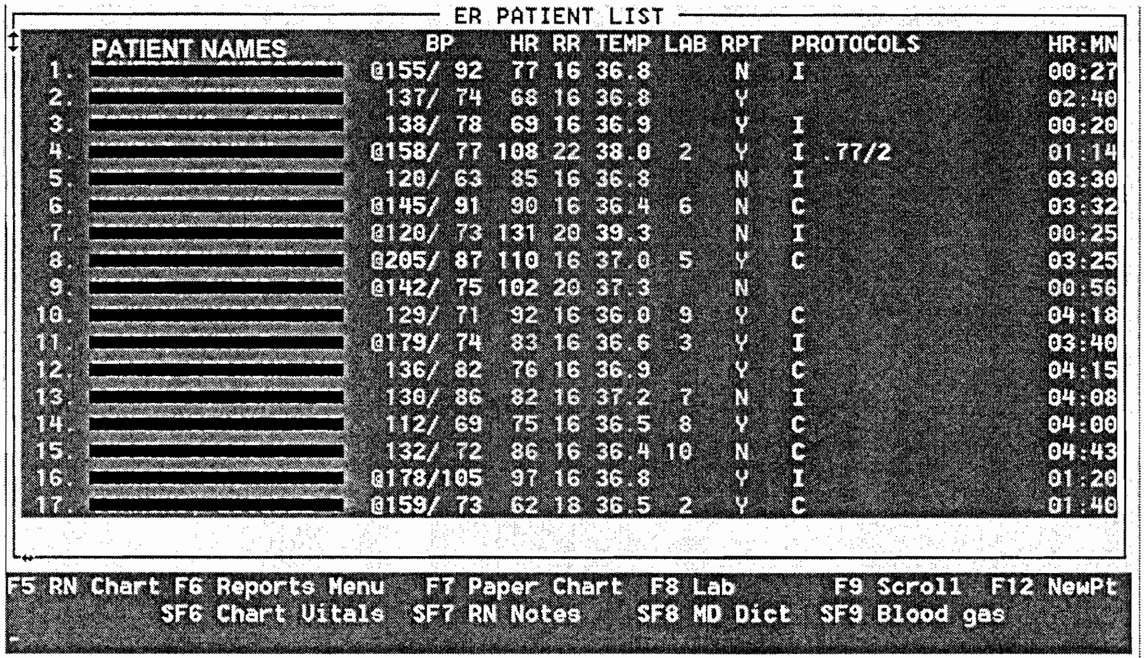
5.8.4 Cross-Over Design

Some of the above mentioned effects may be counterbalanced with a mixed design, or more specifically with a cross-over or split-plot design (27). Compared to a simple randomized design the cross-over design potentially increases the power of the study and may require a smaller sample size. The potential efficiency of the design is based on the within-subject rather than the between-subject observations used in the parallel-group design. The within-subject comparison may be more efficient if considerable variability between subjects exists. However, the crossover design has several drawbacks that may jeopardize the internal validity of the study. In our evaluation a major disadvantage to be considered was the presence of a carry-over effect that could obscure the presence of an effect. Similar to the repeated measures design introducing a washout period is a countermeasure for carry-over effects. Due to the seasonal variation in pneumonia and the time limitation of the evaluation the cross-over design was not our primary choice.

We finally chose a traditional experimental design that randomized patients rather than ED physicians into an intervention and a comparison group (28). Randomizing patients may circumvent problems involved in the physician cross-over or the experimental design with ED physician randomization. To achieve balanced sample sizes for the large number of ED patients during the study period we applied a block randomization with blocks of 12 consecutive ED patients. Each block had six patients in the intervention and six patients in the comparison group. The randomization was performed immediately after registration and before any data elements were available. Because at the time of randomization we did not know the patient's final diagnosis it was not feasible to randomize pneumonia patients into intervention and comparison groups.

The DDSS information was displayed on the ED main screen where anyone with a legitimate log-in was able to observe the information. For the majority of ED patients the pneumonia probability remained below the probability threshold and was not displayed to avoid overloading ED physician with useless data. A character distinguished between patients in the intervention and comparison group (Figure 5.1).

Displaying a patient’s randomization status allowed the ED physicians to recognize whether pneumonia information might become available during a patient's ED encounter. Hiding a patient’s randomization status produces an ambiguous situation and leaves room for two different interpretations. In the first situation, the patient is in the intervention group, but the probability threshold for pneumonia was not crossed. In the second situation, the patient is in the comparison group and no information is displayed.



ER PATIENT LIST

	PATIENT NAMES	BP	HR	RR	TEMP	LAB	RPT	PROTOCOLS	HR:MN
1.	[redacted]	@155/ 92	77	16	36.8		N	I	00:27
2.	[redacted]	137/ 74	68	16	36.8		Y		02:40
3.	[redacted]	138/ 78	69	16	36.9		Y	I	00:20
4.	[redacted]	@158/ 77	108	22	38.0	2	Y	I .77/2	01:14
5.	[redacted]	120/ 63	85	16	36.8		N	I	03:30
6.	[redacted]	@145/ 91	90	16	36.4	6	N	C	03:32
7.	[redacted]	@120/ 73	131	20	39.3		N	I	00:25
8.	[redacted]	@205/ 87	110	16	37.0	5	Y	C	03:25
9.	[redacted]	@142/ 75	102	20	37.3		N		00:56
10.	[redacted]	129/ 71	92	16	36.0	9	Y	C	04:18
11.	[redacted]	@179/ 74	83	16	36.6	3	Y	I	03:40
12.	[redacted]	136/ 82	76	16	36.9		Y	C	04:15
13.	[redacted]	130/ 86	82	16	37.2	7	N	I	04:08
14.	[redacted]	112/ 69	75	16	36.5	8	Y	C	04:00
15.	[redacted]	132/ 72	86	16	36.4	10	N	C	04:43
16.	[redacted]	@178/105	97	16	36.8		Y	I	01:20
17.	[redacted]	@159/ 73	62	18	36.5	2	Y	C	01:40

Figure 5.1: ED main screen. The ED main screen is the most common entry screen for reviewing and charting patient information. The patient names are accompanied by the main vital signs and the number of available laboratory (column "LAB") and dictated hospital reports (column "RPT"). Abnormal vital signs are flagged in the first column ("@") and the respective values are displayed in a different color. The last column ("PROTOCOLS") displays the pneumonia probability and the pneumonia severity index. For the evaluation study an "I" informed users that the patient was assigned to the intervention group in which DDSS information might become available during the patient's encounter. For patients assigned to the comparison group ("C") no information will be available even though the patient might have pneumonia and a high probability. For the patients currently in the ED the pneumonia information is available for one patient of the intervention group and shows a 77 percent probability of pneumonia and stratified the patient into the pneumonia risk class 2.

Displaying the patient's assigned randomization group resolves this ambiguity. For example, absent DDSS information in an intervention group patient tells the ED physician that the probability threshold for pneumonia was not crossed.

We chose the patients rather than the physicians as the unit of analysis and randomization. Choosing the patient as the unit of analysis might violate the independence assumption. The independence assumption requires attention in designs with a nested structure and addresses a possible correlation between the different hierarchical levels (29). The assumption is that establishing a diagnosis is independent of the ED physician and that there are no differences among ED physicians for making the diagnosis. In our evaluation patients are nested in ED physicians, and the diagnostic ability of ED physicians in establishing a diagnosis of pneumonia may vary. Large differences in diagnostic ability among ED physicians may violate the independence assumption and result in an increased likelihood of committing a type I error. However, we did not expect large differences of diagnostic ability among ED physicians. In a situation where the independence assumption does not hold, hierarchical linear modeling techniques are able to account for related effects in nested designs, but were not available until recently (30).

5.9 Discussion

The demand for integrated decision support systems grows as an increasing number of hospitals depend on clinical information systems. To explore new algorithms stand-alone DDSSs will continue to be developed and evaluated in an artificial laboratory setting. However, to effectively support clinicians for routine patient care, decision support systems need to be integrated into clinical information systems and into the physicians' workflow. The evaluation of an integrated system challenges researchers because the characteristics of a clinical setting have an important influence on how the system is applied for the care of patients. The characteristics might be completely independent of the system and might relate to behavioral and psychological issues.

We described the study design for the evaluation of a real-time, integrated decision support system. We illustrated that the clinical setting offers several interesting, but challenging factors. Existing challenges considerably influence the evaluation and are commonly not present in studies that evaluate systems in a more artificial setting. The complexity of a system's evaluation increases as the system moves through the phases of system development to routine clinical application. At higher levels of system implementation the behavioral factors of the targeted users and the logistical aspects of the clinical environment dominate the technical properties of developing and implementing a decision support system. The planning phase is a dynamic process and every study design involves tradeoffs. Some factors may have a considerable influence in one design, but are less influential in another. Many of the described aspects appeared over time during the planning phase. The design of the clinical evaluation for a decision support system remains a challenge, and evaluators have to be flexible enough to balance between feasibility, study design characteristics, statistical considerations, and limited financial and personnel resources.

A clinical evaluation study depends on a fundamental assumption concerning the behavioral aspects of users. The behavioral aspect of users can only be verified during or after the study period: Will ED physicians actually incorporate the unsolicited information into their diagnostic decision process? Evaluators should not take it for granted that a decision support system is welcome and that the information is automatically incorporated in the clinicians' reasoning. If clinicians are required to change processes they are less willing to use the system. Clinicians may incorporate the system's output into their decision making more frequently if the system is highly integrated and delivers information in an unsolicited or easy accessible way.

The eventual purpose of our DDSS is to detect pneumonia patients with high accuracy. If the DDSS accomplishes the diagnostic task, detected patients can be flagged in the clinical information system as pneumonia patients. Based on the pneumonia flag in the clinical information system patient and pneumonia specific protocols or guidelines can be evaluated automatically. Pneumonia guideline candidates that might benefit from a pneumonia flag in the clinical information system include vaccination guidelines (23), criteria for intensive care unit admission (24), or discharge criteria (25).

Because our study evaluated a new approach in a clinical setting we wished to find answers to simple questions. Will clinicians consider the provided information? Will the DSS influence the clinician's diagnostic accuracy? However, questions about changes in behavior and clinical impact on the patients' outcome are eventually of larger interest. Our evaluation study represents an intermediate step in the life cycle of a decision support system. As the system moves through the life cycle, further evaluation studies will be necessary to demonstrate a clinical impact.

5.10 References

1. Miller RA. Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc 1994;1:8-27.

2. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. N Engl J Med 1994;330:1792-6.

3. de Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. Br Med J 1972;2:9-13.

4. Baxt WG, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. Lancet 1996;347:12-5.

5. Aronsky D, Haug PJ. Diagnosing community-acquired pneumonia with a Bayesian network. Proc AMIA Symp. 1998;:632-6.

6. Aronsky D, Haug PJ. An integrated decision support system for diagnosing and managing patients with community-acquired pneumonia. Proc AMIA Symp. 1999;:197-201.

7. Fine MJ, Auble TE, Yealy DM, Hanusa BH, Weissfeld LA, Singer DE, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med 1997;336:243-50.

8. Atlas SJ, Benzer TI, Borowsky LH, Chang Y, Burnham DC, Metlay JP, et al. Safely increasing the proportion of patients with community-acquired pneumonia treated as outpatients: an interventional trial. Arch Intern Med 1998;158:1350-6.

9. Aronsky D, Haug PJ. Assessing the quality of clinical data in a computer-based record for calculating the pneumonia severity index. J Am Med Inform Assoc. 2000;7:55-65.

10. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med. 1997;16:981-91.

11. Poses RM, Cebul RD, Collins M, Fager SS. The importance of disease prevalence in transporting clinical prediction rules. The case of streptococcal pharyngitis. Ann Intern Med 1986;105:586-91.

12. Gallagher EJ. Clinical utility of likelihood ratios. Ann Emerg Med 1998;31:391-7.

13. Metlay JP, Schulz R, Li YH, Singer DE, Marrie TJ, Coley CM, et al. Influence of age on symptoms at presentation in patients with community-acquired pneumonia. Arch Intern Med. 1997;157:1453-9.

14. Whittle J, Fine MJ, Joyce DZ, Lave JR, Young WW, Hough LJ, et al. Community-acquired pneumonia: can it be defined with claims data? Am J Med Qual 1997;12:187-93.

15. Guevara RE, Butler JC, Marston BJ, Plouffe JF, File TM Jr, Breiman RF. Accuracy of ICD-9-CM codes in detecting community-acquired pneumococcal pneumonia for incidence and vaccine efficacy studies. Am J Epidemiol 1999;149:282-9.

16. Wipf JE, Lipsky BA, Hirschmann JV, Boyko EJ, Takasugi J, Peugeot RL, et al. Diagnosing pneumonia by physical examination: relevant or relic? Arch Intern Med 1999;159:1082-7.

17. Albaum MN, Hill LC, Murphy M, Li YH, Fuhrman CR, Britton CA, et al. Interobserver reliability of the chest radiograph in community-acquired pneumonia. PORT Investigators. Chest 1996;110:343-50.

18. Fine MJ, Stone RA, Singer DE, Coley CM, Marrie TJ, Lave JR, et al. Processes and outcomes of care for patients with community-acquired pneumonia: results from the Pneumonia Patient Outcomes Research Team (PORT) cohort study. Arch Intern Med 1999;159:970-80.

19. Corley DE, Kirtland SH, Winterbauer RH, Hammar SP, Dail DH, Bauermeister DE, et al. Reproducibility of the histologic diagnosis of pneumonia among a panel of four pathologists: analysis of a gold standard. Chest 1997;112:458-65.

20. Friedman CP, Wyatt JC. Evaluation Methods in Medical Informatics (Computers and Medicine). New York: Springer-Verlag, 1997.

21. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207-15.

22. Greenes RA, Begg CB. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients. Invest Radiol 1985;20:751-6.

23. Ewig S. Community-acquired pneumonia: definition, epidemiology, and outcome. Semin Respir Infect. 1999;14:94-102.

24. Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Med Inform (Lond) 1990;15:205-17.

25. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29-36.

26. Drew CJ, Hardman ML, Hart AW. Designing and Conducting Research : Inquiry in Education and Social Science. Allyn & Bacon, 1995.

27. Woods JR, Williams JG, Tavel M. The two-period crossover design in medical research. Ann Intern Med 1989;110:560-6.

28. Morris AH, Wallace CJ, Menlove RL, Clemmer TP, Orme JF Jr, Weaver LK, et al. Randomized clinical trial of pressure-controlled inverse ratio ventilation and

extracorporeal CO2 removal for adult respiratory distress syndrome. Am J Respir Crit Care Med 1994;149:295-305.

29. Crits-Christoph P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. J Consult Clin Psychol 1991;59:20-6.

30. Sullivan LM, Dukes KA, Losina E. Tutorial in biostatistics. An introduction to hierarchical linear modeling. Stat Med 1999;18:855-88.

CHAPTER 6

PROSPECTIVE VALIDATION OF A REAL TIME BAYESIAN NETWORK FOR THE AUTOMATIC IDENTIFICATION OF PNEUMONIA PATIENTS

Dominik Aronsky, MD, Peter J Haug, MD

Lancet 2000 (submitted)

6.1 Abstract

Background: Bayesian networks are probabilistic systems that can be applied to diagnostic tasks. The automatic identification of pneumonia patients is such a diagnostic task. Identifying patients in real time using data routinely available in a clinical information system can be used to automatically trigger computerized guidelines or predictive instruments during encounter. We report the prospective clinical validation of a real time Bayesian network for the automatic identification of pneumonia patients.

Methods: A Bayesian network developed from 32,000 emergency department patients from a tertiary care hospital was tested and optimized to identify patients likely to have pneumonia. Only data routinely available in our clinical information system was used. During a 5-month period the Bayesian network was prospectively applied to all emergency department patients 18 years and older. The Bayesian network computed and continuously updated a patient's probability of pneumonia during the encounters of 10,828 patients (265 with pneumonia) in the emergency department. Eight physicians not belonging to the emergency department or the development team established a gold standard for pneumonia. We evaluated the system's diagnostic performance using the area under the receiver operating characteristic curve.

Findings: The Bayesian network included 24 variables, 38 links, and 9,784 conditional probabilities. The area under the receiver operating characteristic curve was 0.942 (95 percent confidence interval: 0.927, 0.955).

Interpretation: The application of an integrated Bayesian network using patient information routinely available in a clinical information system and during a patient's

emergency department encounter appears to be a feasible method for the real time, automatic identification of pneumonia patients.

6.2 Introduction

The clinical application of disease specific predictive instruments and guidelines requires that the patient's disease status be known. Although diagnostic computer systems have a long tradition (1) and have shown remarkable diagnostic performance (2-6), most are not practical for routine patient care because they are typically deployed as stand alone systems that require physicians to perform the time consuming and often redundant task of entering data. Clinical information systems are becoming the standard for capturing, storing, and reporting patient information. Although clinical information systems are rich data sources, most have not been applied for real-time diagnostic tasks. Virtually all computerized diagnostic systems are stand-alone systems that require clinicians to enter the patient information available at that time. Attempts to apply routinely collected and available information from a clinical information system to determine the likelihood of a patient's disease state have not been made. Using routinely available information to automatically identify patients with a specific disease during the encounter might reduce or eliminate data entry by clinicians and avoid compromising the physician's time spent with the patient. This study validated a real time Bayesian network designed to automatically detect patients likely to have pneumonia in an emergency department population using only data routinely available in the clinical information system during the patient's encounter.

6.3 Methods

6.3.1 Patient information

The study was performed at the emergency department (ED) of LDS Hospital (Salt Lake City, Utah), a 425-bed tertiary care hospital. The ED uses a clinical information system for capturing and reviewing patient information. During the 5-month study period, between November 12, 1999, and April 15, 2000, we included all patients 18 years and older with a computerized patient record. Follow-up encounters to the ED were not considered.

6.3.2 Bayesian network

Bayes' theorem for combining evidence and updating beliefs has a long history in medicine (7). As stand-alone tools, simple Bayesian systems were applied to study diagnostic medical problems as early as 1961 (8). BNs are probabilistic models that represent the complex conditional dependencies between clinical findings, symptoms, and diseases (9). Each finding, symptom, or disease is represented as a node, and nodes are connected through links modeling the dependencies among findings, symptoms, and diseases. A table containing the probabilities for each state conditioned on its parents is attached to the nodes. These conditional probabilities are used for revising the joint probability of a BN by applying Bayes' theorem. The conditional probability distributions for each node in the BN are estimated by domain experts or trained from large data sets. BNs account for missing information by using prior probability distributions when updating probabilities. In addition, BNs can be examined in detail and used to determine the contribution of each finding and symptom to the disease probability.

6.3.3 Design

Objective – Physicians remain responsible to identify patients eligible for a guideline, a manual task that limits guideline implementation. The goal of the BN was to automatically identify ED patients likely to have pneumonia. To realize complete automation the BN had to operate in real time and with no or minimal additional data entry from health care providers. For this reason we included only data variables routinely available during the patient's ED encounter.

Development - To develop the BN we created a data set from our clinical information system. The data set included 72 variables from more than 32,000 ED patients at LDS hospital during a 17-month period (June 1996 – November 1997) and included 498 pneumonia patients identified by primary ICD-9 discharge diagnoses (32.8, 480 – 486; 518.81 or 518.82 with a pneumonia code in secondary position). We developed and evaluated more than 100 different BNs using the software Netica™ from Norsys®. We evaluated the different BNs with a 3-leave-1-out jackknife method. For the 3 runs we calculated the average area under the receiver operating characteristic curve (ROC) to determine diagnostic accuracy (10). Between the time of BN development and clinical validation the ED clinical information system underwent minor documentation changes, such as revising the list of codable chief complaints. Because minor changes in practice are typical for a clinical setting we did not attempt to account for the changes in the BN knowing that diagnostic accuracy might slightly suffer.

Prospective validation - For every patient in the ED the BN queried the clinical information system for new data every 5 minutes during the patient's encounter. If new patient information became available, the BN updated the pneumonia probability to reflect all available information at that moment. Probability updating concluded at patient discharge yielding the final pneumonia probability for the evaluation. The BN was operational 98.9 percent during the study period; the system was inactive when the hospital network or the clinical information system was down. Both the Internal Review Boards of the study site hospital and the university approved the study.

6.3.4 Final diagnosis

There are no objective criteria for the diagnosis of pneumonia. To avoid verification bias (11) the disease status of all ED patients was considered with respect to the presence of pneumonia. From all ED study patients we selected all patients who met at least one of the following criteria:

(1) patients with a chest radiology exam (chest x-ray or tomogram) performed during the ED encounter;

(2) patients with pneumonia compatible chief complaints where a pneumonia compatible chief complaint was defined as any chief complaint that a pneumonia patient complained of in the 32,000 patient development set;

(3) patients with a primary admit or discharge ICD-9 diagnosis of pneumonia;

(4) patients with a BN calculated probability of 1 percent or higher;

(5) patients with the term "pneumonia" in any dictated report (ED report, admission report, ED follow-up report, operation report, discharge summary).

If a patient did not meet any of the 5 criteria we considered the chances of pneumonia as extremely low and, consequently, pneumonia to be absent in such patients.

For patients meeting at least 1 of these criteria, 5 internal medicine physicians (1 second-year resident, 2 third-year residents, and 2 board-certified internists) read the patients' ED reports and chest exam reports. Based on the reports the 5 review physicians were asked to select patients without any evidence for pneumonia and to exclude these patients from further review.

For patients determined to have a remaining chance of pneumonia, 3 physicians, board-certified in pulmonary and critical care medicine, reviewed the patients' medical charts and chest x-rays. A radiological confirmed infiltrate within 48 hours of presentation and the presence of at least one finding suggestive of pneumonia (e.g. fever, shortness of breath, chest pain, cough, dyspnea, etc.) was required to establish a pneumonia diagnosis (12). The majority vote of the 3 reviewers determined the diagnosis of pneumonia. Majority vote was established by involving the third reviewer in cases where the other two reviewers disagreed.

In patients determined to have pneumonia, reviewers assessed the types of pneumonia, which included community-acquired (including nursing home patients), hospital-acquired pneumonia, pneumonia in immuno-compromised patients, pneumonia in patients with evidence of Mycobacterium tuberculosis infection, and postobstructive pneumonia in bronchial malignancy (13). Additionally, reviewers determined the presence of aspiration pneumonia, but this diagnosis was not considered as pneumonia for the purpose of this study. No emergency department physician and no member of the BN development team was a physician reviewer.

6.3.5 Statistical Analysis

An ROC curve was created and the area under the ROC with 95 percent confidence intervals was computed using maximum likelihood estimation (14). The area under the ROC curve is an overall accuracy measure for the predictive power of an instrument (10). As with any probabilistic predictive instrument distinct cutoff points can be chosen depending on the instrument's intended purpose. Because the BN's purpose is to automatically identify patients likely to have pneumonia and automatically trigger computerized guideline evaluation, we favored high sensitivity over high specificity. At 95 percent and 90 percent sensitivity levels we determined the respective specificity, positive and negative predictive values, and odds likelihood ratio, including 95 percent confidence intervals where appropriate. We applied the likelihood ratios to calculate a standardized test effectiveness statistic δ (6, 15):

$$\delta = (\sqrt{3} / \pi) * (\ln [\text{sensitivity} / 1\text{-specificity}] + \\ \ln [\text{specificity} / 1\text{-sensitivity}])$$

The statistic δ is a measure of discriminatory power of a test and sums the log of the positive and the negative likelihood ratios scaled by the standard deviation of the logistic normal distribution ($\sqrt{3} / \pi$).

6.4 Results

The final, most parsimonious model included 24 variables, 38 conditional links, and 9,784 conditional probabilities (Figure 6.1). The BN computed a pneumonia probability for 10,828 patients of whom 265 patients had a gold standard diagnosis of pneumonia.

Input variables

Patient information	Age	cont
	Number of risk factors	cont
Vital signs	Heart rate	cont
	Respiratory rate	cont
	Systolic blood pressure	cont
	Oxygen saturation	cont
	Temperature	cont
Laboratory results	White blood count	cont
	Bands	cont
	Blood urea nitrogen	cont
	γ-glutamyl transferase	cont
	Alkalic phospatase	cont
Blood gas results	Arterial O_2 pressure	cont
	Arterial CO_2 pressure	cont
	pH	cont
Nurse Assessment	Chief complaint	cat
	Cough	cat
	Breath sounds	cat
	Abdominal exam	cat
	Abdominal discomfort	cat
	Pain characteristic	cat
Chest x-ray	Order	dich
	Interpretation	dich

Optimal Bayesian network structure (output: pneumonia probability)

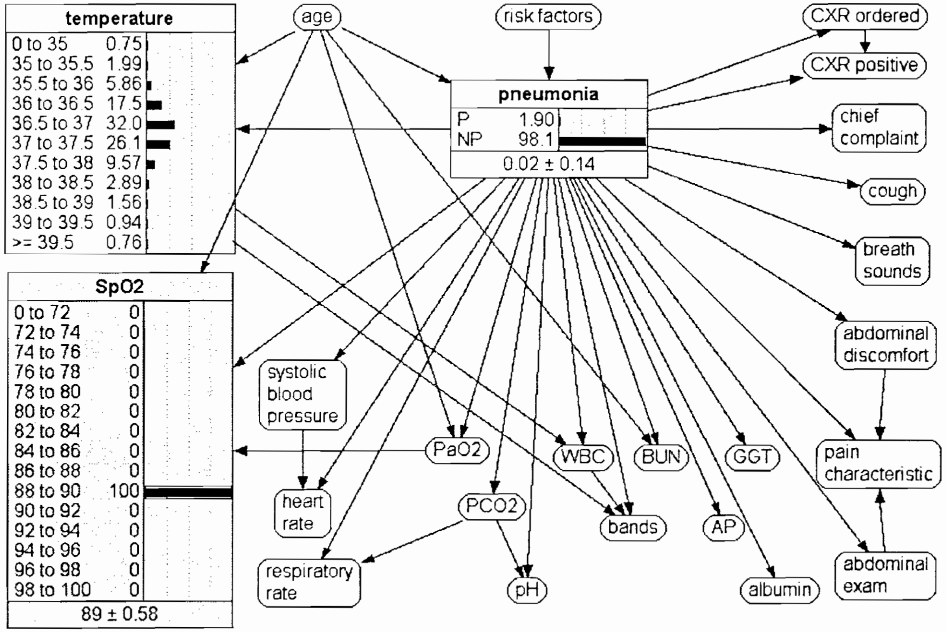


Figure 6.1: Most parsimonious Bayesian network structure. The node "SpO_2" is instantiated reflecting the known evidence of oxygen saturation, the node "temperature" shows the prior probability distribution for all patients with the known oxygen saturation value, and "pneumonia" reflects the current belief for the disease state. Cont = continuous; cat = categorical; dich = dichotomous.

The types of pneumonia were: 236 community-acquired, 15 hospital-acquired, 13 pneumonia in immuno-suppressed patients, and 1 hospital-acquired in an immuno-suppressed patient. The pneumonia prevalence during the study period was 2.4 percent.

Figure 6.2 shows the ROC curve for the prospective validation set. The area under the ROC curve was 0.942 (95 percent confidence interval: 0.927; 0.955). At a 90 percent sensitivity level the BN's specificity, positive and negative predictive values, and positive and negative likelihood ratios were 0.853 (95 percent confidence interval: 0.860; 0.846), 0.133 (0.091; 0.175), 0.9970 (0.9959; 0.9981), 6.10 and 0.12. At a 95 percent sensitivity level the BN's specificity, positive and negative predictive values, positive and negative likelihood ratios were 0.774 (95 percent confidence interval: 0.765; 0.783), 0.095 (0.06; 0.131), 0.9984 (0.9976; 0.9993), 4.20 and 0.06. Figure 6.3 illustrates the test effectiveness statistic of the BN over the range of sensitivity/specificity values. The test effectiveness statistic was 2.17 at the 90 percent and 2.31 at the 95 percent sensitivity level. The results are summarized in Table 6.1.

6.5 Discussion

Diagnosing diseases with computer support has attracted investigators for a long time. Diagnostic systems have been developed that cover a broad range of diseases (2) or focus on a smaller, more focused area of medicine (3-6). These diagnostic systems, however, are not used for routine patient care. One reason is that these systems are stand-alone and lack integration with clinical information systems. With stand-alone systems clinicians are required to enter data and spend time interacting with the computer rather than the patient. Clinical information systems are becoming widely available and

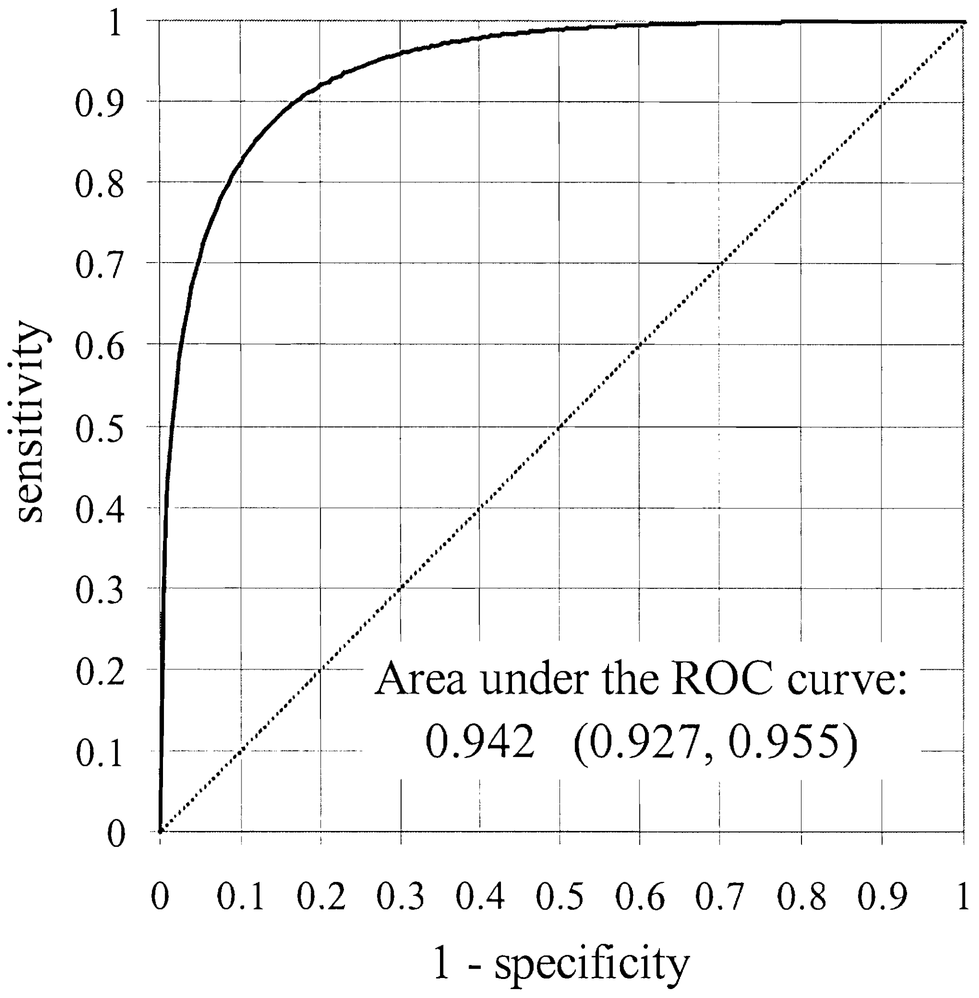


Figure 6.2: ROC curve for the Bayesian network. A random generator produces an AUC of 0.5 (dotted line) and a perfect predictive instrument an AUC of 1.0.

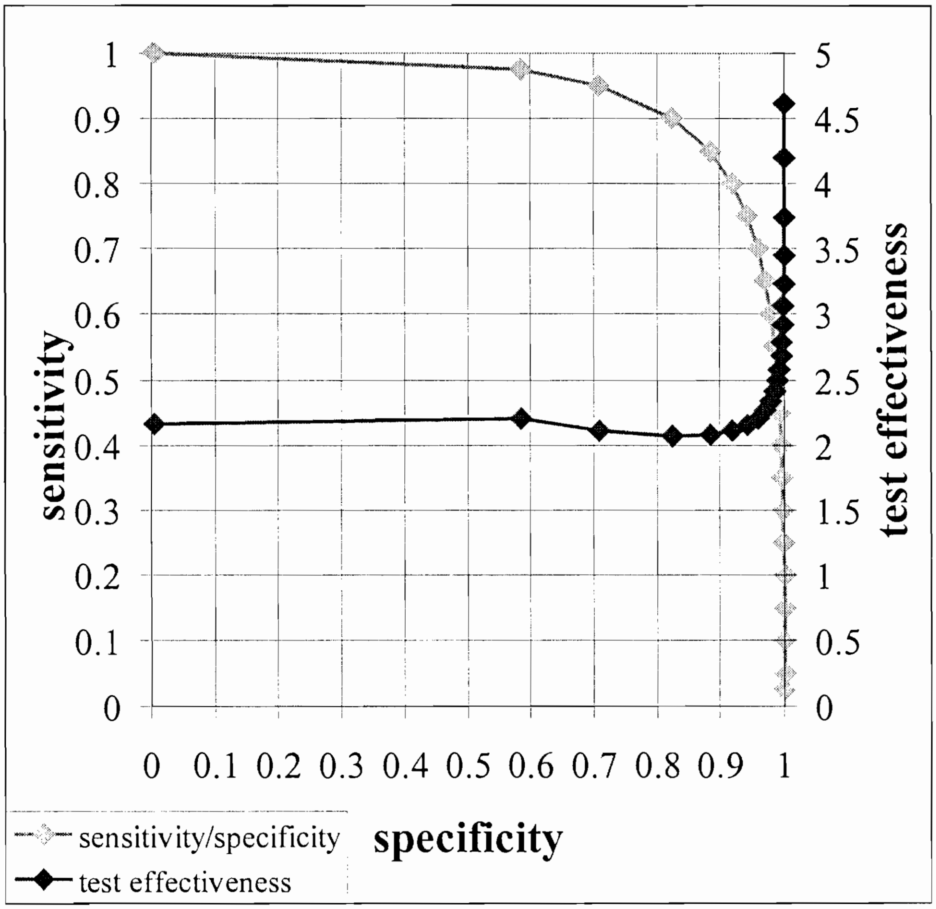


Figure 6.3: Test effectiveness of the Bayesian network plotted against respective pairs of sensitivity and specificity. Test effectiveness shows stable discriminatory power with high sensitivity values. The relatively low pneumonia prevalence is a result of the population based study and yields heavily unbalanced contingency tables. That is why the discriminatory power increases only beyond specificity of 0.99.

Table 6.1: Performance Characteristics of Bayesian Network (at fixed sensitivity levels)

Sensitivity	Specificity	Positive predictive value	Negative predictive value	Positive likelihood ratio	Negative likelihood ratio	Test effectiveness statistic
0.90	0.853	0.133	0.9970	6.10	0.12	2.17
0.95	0.774	0.095	0.9984	4.20	0.06	2.31

routinely used for capturing and storing patient information. Linking decision support systems to clinical information systems has demonstrated measurable value in patient care (16). Decision support systems with a diagnostic task, however, have not been integrated with clinical information systems.

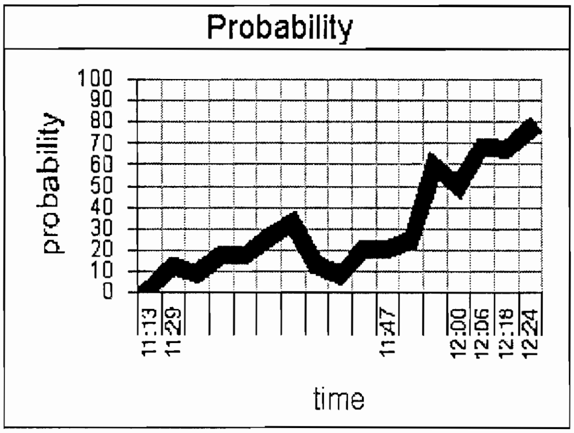
The diagnostic performance of stand-alone systems has demonstrated considerable accuracy (2-6) and, on occasion, outperformed clinicians (6). The AUC of growing cell structure networks applied to the cytological diagnosis of breast carcinoma was 0.96 (17). In a comparative study for the diagnosis of acute cardiac ischemia the AUC of logistic regression was 0.905, of a classification tree model 0.861, and an artificial neural network 0.923 (18). The test effectiveness statistic of a prospectively validated artificial neural network for the diagnosis of patients with anterior chest pain was 3.50 at a 96 percent sensitivity level (6). Our BN had an AUC of 0.942 and a test effectiveness statistic of 2.31 at the 95 percent sensitivity level. Our system operated on an entire emergency department population, and the prevalence of the target disease was low (2.4 percent). We did not restrict the BN's application to patients meeting specific chief complaints or clinical findings. In situations of low disease prevalence it becomes challenging to achieve high diagnostic accuracy. The BN functioned without requiring health care providers to enter additional data. This reduces the clinicians' time spent interacting with the computer. In view of the BN's characteristics we consider the diagnostic accuracy to be high and, thus, the BN may represent a practical and valuable tool for the automatic identification of patients with pneumonia.

Our goal was to develop an application suitable for clinical use and to test whether data available from routine patient documentation can be utilized for a computerized diagnostic task. The goal was not to compete with the diagnostic expertise of physicians, but instead to detect patients with a disease for whom disease specific guidelines and predictive instruments exist, but are not applied in clinical settings.

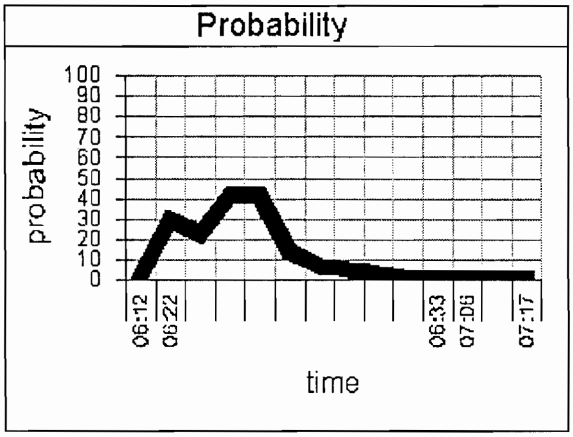
A common objection of clinicians is the "black box" behavior of artificial neural networks, which cannot to give clinicians insight into the reasoning process. BNs can show clinicians a trace of their probabilistic reasoning (Figure 6.4). This may represent an advantage over other statistical representations. However, it remains unknown whether the clinical application of mathematical models is more successful when the system can offer explanations.

There are limitations to our study. For outpatients clinical information is limited to the data available during the ED encounter. It is conceivable that outpatients diagnosed with a respiratory illness other than pneumonia failed current treatment and had a follow-up visit at a different medical facility where they were diagnosed with pneumonia. We were unable to follow up on such patients. The quality of data captured in the clinical information system during a patient's clinical encounter may be questioned. The majority of patient information in clinical information systems and in our system originates from the nurses taking care of patients. Clinicians might challenge the accuracy of patient information when such data are used for diagnostic purposes. However, these were the data available in our information system, and they allowed us to automate the identification process. We can only guess what the system's diagnostic performance would be if physicians entered patient data in the information system at the point of care.

The development and the prospective validation have been performed at a single institution. How and what type of patient data are charted in the clinical information



Time	Variable	Value	Probability
11:13	Prior probability		.016
11:29	Chief complaint	Respiratory	.124
11:29	Current history	Risk factors	.088
11:29	Past history	Risk factors	.17
11:29	Current medication	Risk factors	.17
11:29	Heart rate	108	.27
11:29	Systolic BP	158	.132
11:29	Respiratory rate	22	.324
11:29	Temperature	38.0	.207
11:29	Oxygen saturation	92	.082
11:47	Abdominal exam	Not distended	.198
11:47	Cough	Non-productive	.243
11:47	Breath sounds	Rales	.59
12:00	CXR order	Chest 2 Views	.496
12:06	White blood count	21.4	.684
12:18	Bands	3	.665
12:24	Blood urea nitrogen	25	.776



Time	Variable	Value	Probability
06:12	Prior probability		.016
06:22	Chief complaint	Respiratory	.296
06:22	Current history	Risk factors	.227
06:22	Past history	Risk factors	.418
06:22	Current medication	Risk factors	.418
06:22	Heart rate	66	.139
06:22	Respiratory rate	16	.068
06:22	Systolic BP	140	.051
06:22	Oxygen saturation	90	.031
06:22	Temperature	35.5	.005
06:33	Abdominal exam	Not distended	.005
07:06	White blood count	12.4	.003
07:06	CXR order	Chest 1 View	.002
07:17	Blood urea nitrogen	39	.001

Figure 6.4: Probabilistic reasoning of the Bayesian network: Course of probability during the encounter of a 77-year old patient with pneumonia and a 90-year old patient without pneumonia. The temporal availability, the clinical variable including its value and the resulting probability changes show the BN’s ability to offer insight into the probabilistic reasoning. The BN allows a detailed examination of the impact of each clinical variable on the pneumonia probability.

system is customized to a location. The site-specific characteristics of a clinical information system might influence the BN's behavior, and the BN's transportability to other sites remains to be demonstrated.

Successful identification of patients with a real time diagnostic system can support guideline implementation efforts. Identifying patients eligible for a guideline remains a manual task involving the physician who wishes to apply a disease specific guideline. The placement of a pneumonia flag in the patient's electronic record is another application of the automatic disease identification process. Usually the patient's diagnosis is not accessible in computable format until after the patient's discharge because the diagnosis remains concealed in free text reports, problem lists, or hand written progress notes (19). Marking a patient's electronic record can be used to initiate pneumonia guidelines and predictive instruments, such as the Pneumonia Severity of Illness Index (12), pneumonia vaccination guidelines, or discharge criteria. Indicating a patient's disease in the clinical information systems in computable and decidable format might ease and support the implementation of a variety of tools during the encounter of hospitalized patients.

In conclusion, our data suggest that an integrated, real time BN can be, with a high level of accuracy, an effective instrument for automatically identifying patients likely to have pneumonia. The BN can be applied to initiate the evaluation of computerized guidelines and predictive instruments, a manual task commonly performed only by interested clinicians.

6.6 References

1. Miller RA. Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc. 1994;1:8-27.

2. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994;330:1792-6.

3. Pozen MW, D'Agostino RB, Selker HP, Sytkowski PA, Hood WB Jr. A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. A prospective multicenter clinical trial. N Engl J Med. 1984;310:1273-8.

4. Selker HP, Beshansky JR, Griffith JL, Aufderheide TP, Ballin DS, Bernard SA, et al. Use of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms suggestive of acute cardiac ischemia. A multicenter, controlled clinical trial. Ann Intern Med. 1998;129:845-55.

5. deDombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. Br Med J. 1972;2:9-13.

6. Baxt WG, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. Lancet. 1996 Jan 6;347:12-5.

7. Gardner RM, Pryor TA, Warner HR. The HELP hospital information system: update 1998. Int J Med Inf. 1999;54:169-82.

8. Ledley RS, Lusted LB, Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Scienc 1959;130:9-21.

9. Warner HR, Toronto AF, Veasey LG, Stephenson RA. Mathematical approach to medical diagnosis. JAMA. 1961;177:75-81.

10. Jensen FV. An introduction to Bayesian Networks, Springer-Verlag New York Inc., New York, 1997.

11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29-36.

12. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983;39:207-15.

13. Fine MJ, Auble TE, Yealy DM, Hanusa BH, Weissfeld LA, Singer DE, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. N Engl J Med. 1997;336:243-50.

14. Ewig S. Community-acquired pneumonia: definition, epidemiology, and outcome. Semin Respir Infect. 1999;14:94-102.

15. ROCKIT (0.9B Beta Version). Charles E. Metz. Dept. of Radiology, Univ. of Chicago.

16. Blakeley DD, Oddone EZ, Hasselblad V, Simel DL, Matchar DB. Noninvasive carotid artery testing. A meta-analytic review. Ann Intern Med.1995;122:360-7.

17. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA. 1998;280:1339-46.

18. Walker AJ, Cross SS, Harrison RF. Visualisation of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique. Lancet. 1999;354:1518-21.

19. Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. J Investig Med. 1995;43:468-76.

20. Tierney WM, Overhage JM, McDonald CJ. Toward electronic medical records that improve care. Ann Intern Med 1995;122:725-6.

CHAPTER 7

DISCUSSION

7.1 Discussion

The objective of this project was to overcome the “behavioral bottleneck” that exists when clinicians are required to identify guideline eligible patients, initiate computerized guideline evaluation, and enter patient-specific information. In an attempt to master the “behavioral bottleneck” I developed a decision support system that was required to meet the following specifications: (1) identify pneumonia patients with high accuracy; (2) function in real time to allow the delivery of information during a patient’s encounter; (3) operate without requiring health care providers to enter additional data.

The development and implementation of decision support systems occurs in phases (1, 2). The first step involves the development and testing of a model in a laboratory setting. The last step involves clinical studies that demonstrate added value for health care, change in patient outcomes, and improved patient care. Optimally, an effect should be demonstrated in multicenter studies showing the system’s transportability to other sites.

A system that goes through all the phases will likely require several years of development, implementation, and evaluation. At any phase researchers should be prepared to go back to an earlier phase and make system modifications that are needed for successful implementation in a subsequent phase. In software engineering this process is a common task necessary to design a successful, customer-oriented product. Releasing a new version of software is similar to a clinical study with the exception that the cycle of product improvement is several times faster for a commercial software product than for a medical decision support system. This might be one of several reasons why many promising decision support systems do not reach the stage of clinical evaluation. For the clinical evaluation of decision support systems there are four fundamental questions to be addressed (3):

(1) Will physicians use the system in a clinical setting?

(2) Will the provided information alter decisions?

(3) Will altered decisions result in a change of behavior?

(4) Will the change of behavior lead to a change of patient outcomes?

From a high level view, the decision support system described in this dissertation has to be considered as an early phase project that demonstrated the feasibility of the chosen approach. I showed that a diagnostic algorithm could be used to circumvent the "behavioral bottleneck" from a medical informatics viewpoint. The system could detect patients eligible for a disease specific guideline. The automatic identification process was used to initiate computerized guideline evaluation and provide clinicians with the respective recommendations in an unsolicited way. Although the chosen approach of detecting patients automatically appears to be feasible, it does not guarantee that we can master the "behavioral bottleneck" from a psychological viewpoint.

Delivering information in an automatic and unsolicited way does not necessarily mean that physicians will consider the information for patient care (4). In parallel with the clinical evaluation of the diagnostic component I examined whether displaying the pneumonia probability and the Pneumonia Severity of Illness risk class in an automatic and unsolicited way had an influence on the admission behavior of emergency physician. The prospective study randomized emergency department patients to an intervention and a comparison group. For patients in the intervention group pneumonia information was displayed while no information was available for patients in the comparison group. The analyses showed no differences in admission behavior and suggest that the physicians might have seen the information, but did not apply computerized pneumonia data for patient care. It remains unknown whether physicians resisted applying the severity index as a clinical risk assessment instrument or whether the computerized approach of automatic information delivery was ineffective.

The function of the diagnostic system was not designed to challenge the physicians' diagnostic skills, but rather to support the computerized implementation and dissemination of the Pneumonia Severity of Illness Index. I chose the Pneumonia Severity of Illness Index as one representative tool among several available clinical guidelines or predictive instruments that were available for the care of pneumonia patients (5-8). There were both clinical and medical informatics reasons for choosing the Pneumonia Severity of Illness Index. The clinical reason was that the severity index is a well-studied tool meeting high quality recommendations for the development of predictive instruments. The medical informatics reason was that the severity index uses only variables that are routinely available during a patient's encounter and are, with one exception, available in computerized form in our clinical information system.

The variable not available in computerized form, the presence or absence of a "pleural effusion," originated from the radiologist's chest x-ray interpretation, which was available in free text format only. The diagnostic system included variables that were available in coded and numerical format. Short free text phrases were also included, but the system did not have the ability to process information from free text reports, such as chest x-ray reports. Many pieces of clinical information are locked in free text reports and difficult to access and use for decision support systems. Natural language processing and understanding methods could contribute to make information from free text reports available for computerized decision support systems.

During the clinical evaluation of the diagnostic system a simultaneous, prospective study was performed that investigated the performance of a natural language understanding system for the real time, automatic identification of pneumonia patients from chest x-ray reports (9, 10). For emergency department patients whose chest x-ray reports became available during the encounter, the radiologist's dictated report was retrieved and submitted to the natural language processing system. A logical and interesting future project would consist of combining the pneumonia decision support system with a natural language processing system and testing whether the combination resulted in higher accuracy and more complete evaluation of the severity index. Alternatively, the radiologists' interpretation of chest x-rays often suggests that the radiology findings require clinical correlation. The information from the diagnostic decision support system and the natural language processing system might provide radiologists with clinical information in real time.

Blending several tools together will allow researchers in medical informatics to better use information stored in clinical information systems. Clinical information systems can capture, store, and display patient information in a variety of formats, such as digital images, free text reports, numerical data, coded data, analog signals, or digitized voice. Combining and exploiting these data for creating more powerful and more sophisticated decision support systems will be an exciting but challenging area of research in medical informatics.

7.2 References

1. Stead WW, Haynes RB, Fuller S, Friedman CP, Travis LE, Beck JR, et al. Designing medical informatics research and library--resource projects to increase what is learned. J Am Med Inform Assoc. 1994;1:28-33.

2. Stead WW. Matching the level of evaluation to a project's stage of development. J Am Med Inform Assoc. 1996;3:92-4.

3. Wyatt J, Spiegelhalter D. Field trials of medical decision-aids: potential problems and solutions. Proc Annu Symp Comput Appl Med Care. 1991;:3-7.

4. Lee TH, Pearson SD, Johnson PA, Garcia TB, Weisberg MC, Guadagnoli E, et al. Failure of information as an intervention to modify clinical management. A time-series trial in patients with acute chest pain. Ann Intern Med. 1995;122:434-7.

5. British Thoracic Society: Community-acquired pneumonia in adults in British hospitals in 1982-1983: A survey of aetiology, mortality, prognostic factors, and outcome. Q J Med.1987;62:195-222.

6. Niederman MS, Bass JB, Campbell GD, Fein AM, Grossman RF, Mandell LA, et al. Guidelines for the initial management of adults with community-acquired pneumonia: diagnosis assessment of severity, and initial antimicrobial therapy. Am Rev Respir Dis 1993;148:1418-26.

7. Ewig S, Kleinfeld T, Bauer T, Seifert K, Schafer H, Goke N. Comparative validation of prognostic rules for community-acquired pneumonia in an elderly population. Eur Respir J. 1999;14:370-5.

8. Dean NC, Suchyta MR, Bateman KA, Aronsky D, Hadlock CJ. Implementation of admission decision support for community-acquired pneumonia. Chest. 2000;117:1368-77.

9. Fiszman M, Haug PJ. Using Medical Language Processing to Support Real-Time Evaluation of Pneumonia Guidelines. Proc AMIA Symp. 2000;:235-9

10. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic Detection of Acute Bacterial Pneumonia from Chest X-Ray Reports. J Am Med Inform Assoc. 2000;7:593-604.