Editorial Comments JAMIA

Matching the Level of Evaluation to a Project's Stage of Development

The first issue of the *Journal of the American Medical Informatics Association* (*JAMIA*) included a paper that proposed a framework for design of applied medical informatics projects and the evaluative studies that are part of them.[1] A fundamental concept was that innovation in medical informatics takes place through sequential stages of system development. The first stage is specification, followed by component development, combination of components into a system, integration of the system into the environment, and, finally, routine use. The level of evaluation should be tuned to the stage of development. For example, problem definition and bench testing are appropriate during the specification stage. Field and validity testing become possible as components are combined into a system. Evaluation of efficacy is not practical until the sytem is integrated into the environment.

As a follow-up to this initial paper, I have re-read the 39 manuscripts describing original investigations or research that have been published in the first two volumes of *JAMIA*. I classified each according to the developmental stage of the work being reported and the level of evaluation employed. One goal was to get a feel for how well reports that had passed peer review fit into the proposed framework. Since the initial paper did not include examples, a second goal was to provide now an index for representative work based on the stage of development and level of evaluation. The brief citations that follow identify each article by the volume of *JAMIA* in which it appeared and its starting page number [see the box for author(s) and title].

Fifteen papers reported work in the specification stage of development. Several strategies were used for problem-definition-level evaluations. Test sets of patient records [1:61, 395, 404] and a sample guideline [2:238] were used to categorize the information that must be represented or located to perform a task. Non–case-based examples [2:4, 323] were used to explain models for representing clinical data. Surveys [1:381, 2:374] were utilized to identify computer literacy and attitudes. The bench-testing level of evaluation was also utilized at the specification stage of development. One or more cases [1:218, 249, 2:19, 116] were used to test a model for representing clinical data. A secondary analysis of cases [2:160] was used to demonstrate the utility of a framework for characterizing system use [2:160]. One guideline was incorporated into a clinical system [2:316] to explore the problems of such an application. Two sample implementations were used to demonstrate the feasibility of a framework for representing information sources [2:383].

Eight papers reported work in the component-development stage of development. Each of these involved the bench testing level of evaluation. The use of natural-language processors to map clinical text into a structured database [1:142, 161] was tested through comparison of recall and precision of queries of a resultant demonstration database with those of experts using the source text. The adequacy of a schema for representing data from chest x-ray reports was tested via manual verification of structured data from one of these processors against a subset of the source documents [1:233]. A set of standard terms for representation of nursing care was pilot-tested by showing consistency among terms selected by three coders [1:175]. A test set of cases was used to evaluate variation in model factors [1:272]. A search strategy was tested against manual review of a subset of journals [1:447]. A speech interface was evaluated by

contrasting two grammars [2:36] and by comparison of the recognizer with a hidden experimenter [2:46].

Ten papers reported work at the stage of development in which components were combined into a system. The majority involved the bench-testing level of evaluation. Performances of expert search [1:51] and diagnostic [1:127] systems with test cases were compared with those of human experts. Test cases were used to test multiple search interfaces against each other [1:285] and to compare content captured through a computer interface with that captured in a manual note [2:365]. Databases of case abstracts were used to train and test systems to predict risk

1:61 *Henry, Holzemer, Reilly, Campbell,* Terms Used by Nurses to Describe Patient Problems: Can SNOMED III Represent Nursing Concepts in the Patient Record?

1:395 *Giuse, Huber, Giuse, Brown, Bankowitz, Hunt,* Information Needs of Health Care Professionals in an AIDS Outpatient Clinic as Determined by Chart Review

1:404 *Bates, O'Neil, Boyle, et al.,* Potential Identifiability and Preventability of Adverse Events Using Information Systems

2:238 *Miller, Frawley,* Trade-offs in Producing Patient-specific Recommendations from a Computer-based Clinical Guideline: A Case Study

2:4 *Friedman, Huff, Hersh, Pattison-Gordon, Cimino,* The Canon Group's Effort: Working toward a Merged Model

2:323 *Dolin,* Modeling the Temporal Complexities of Symptoms

1:381 *Brown, Coney,* Changes in Physicians' Computer Anxiety and Attitudes Related to Clinical Information System Use

2:374 *Lang,* Trends in Students' Knowledge, Opinions, and Experience Regarding Dental Informatics and Computer Applications

1:218 *Campbell, Das, Musen,* A Logical Foundation for Representation of Clinical Data

1:249 *Bell, Pattison-Gordon, Greenes,* Experiments in Concept Modeling for Radiographic Image Reports

2:19 *Rector, Glowinski, Nowlan, Rossi-Mori,* Medical-concept Models and Medical Records: An Approach Based on GALEN and PEN&PAD

2:116 *Huff, Rocha, Bray, Warner, Haug,* An Event Model of Medical Information Representation

2:160 *Brennan,* Characterizing the Use of Health Care Services Delivered via Computer Networks

2:316 *Tierney, Overhage, Takesue, et al.,* Computerizing Guidelines to Improve Care and Patient Outcomes: The Example of Heart Failure

2:383 *Patrick, Springer, Mitchell, Sievert,* Virtual Shelves in a Digital Library: A Framework for Access to Networked Information Sources

1:142 *Sager, Lyman, Bucknall, Nhan, Tick,* Natural Language Processing and the Representation of Clinical Data

1:161 *Friedman, Alderson, Austin, Cimino, Johnson,* A General Natural-language Text Processor for Clinical Radiology

1:233 *Friedman, Cimino, Johnson,* A Schema for Representing Medical Language Applied to Clinical Radiology

1:175 *Ozbolt, Fruchtnicht, Hayden,* Toward Data Standards for Clinical Nursing Information

1:272 *Eisenstein, Alemi,* An Evaluation of Factors Influencing Bayesian Learning Systems

1:447 *Haynes, Wilczynski, McKibbon, Walker, Sinclair,* Developing Optimal Search Strategies for Detecting Clinically Sound Studies in MEDLINE

2:36 *Shiffman, Detmer, Lane, Fagan,* A Continuous-speech Interface to a Decision Support System: I. Techniques to Accommodate for Misrecognized Input

2:46 *Detmer, Shiffman, Wyatt, Friedman, Lane, Fagan,* A Continuous-speech Interface to a Decision Support System: II. An Evaluation Using a Wizard-of-Oz Experimental Paradigm

1:51 *Hersh, Hickam, Haynes, McKibbon,* A Performance and Failure Analysis of a SAPHIRE with a MEDLINE Test Collection

1:127 *Long, Naimi, Criscitiello,* Evaluation of a New Method for Cardiovascular Reasoning

1:285 *Haynes, Walker, McKibbon, Johnston, Willan,* Performances of 27 MEDLINE Systems Tested by Searches with Clinical Questions

2:365 *Moorman, van Ginneken, Siersema, van der Lei, van Bemmel,* Evaluation of Reporting Based on Descriptional Knowledge

1:439 *Woolery, Grzymala-Busse,* Machine Learning for an Expert System to Predict Preterm Birth Risk

1:459 *Lowell, Davis,* Predicting Length of Stay for Psychiatric Diagnosis-related Groups Using Neural Networks

2:220 *McDaniel,* Discrete-event Simulation of a Wide-area Health Care Network

1:186 *Chueh, Barnett,* Client–Server Distributed Database Strategies in a Health Care Record System for a Homeless Population

1:339 *Nelson, Gardner, Hedrick, Gould,* Computerized Decision Support for Concurrent Utilization Review Using the HELP System

1:35 *Cimino, Clayton, Hripcsak, Johnson,* Knowledge-based Approaches to the Maintenance of a Large Controlled Medical Terminology

2:102 *Miller, Frawley, Wright, Roderer, Powsner,* Lessons Learned from a Pilot Implementation of the UMLS Information Sources Map

2:307 *Balas, Stockham, Mitchell, Austin, West, Ewigman,* The Columbia Registry of Information and Utilization Management Trials

2:297 *Giuse, Giuse, Miller,* Evaluation of a Long-term Maintenance of a Large Medical Knowledge Base

1:428 *Gardner, Lundsgaarde,* Evaluation of User Acceptance of a Clinical Expert System

2:58 *Shea, Sideli, DuMouchel, Pulver, Arons, Clayton,* Computer-generated Informational Messages Directed to Physicians: Effect on Length of Hospital Stay

[1:439] and length of stay [1:459]. A simulator was demonstrated by modeling two solutions for a test environment [2:220]. The report of the features of a client-server clincial record system together with benchmarks of performance [1:186] represents a field-test level of evaluation of work at this stage of development. Evaluation of computerized concurrent review with manual review in a crossed, blocked design is an example of a validation-level evaluation [1:339]. The evaluation of the utility of database assistance through a longitudinal study of two cohorts is an example of an efficacy-level evaluation.

Six papers reported systems that were at the stage of routine use. The majority had the character of the field-trial level of evaluation [1:35, 2:102, 307] in that they reported what worked and lessons learned. A comparison of knowledge-base disease profiles before and after update [2:297] is a validation-level evaluation. Another assessed impact through a survey to assess system utility [1:428] as perceived by different user groups. A randomized trial of the effect of providing a length-of-stay reminder is an example of an efficacy-level evaluation for work at this stage [2:58].

This placement of papers into categories represents one person's opinion and it is not precise. Most papers fit into the framework cleanly. An exception involved a case [2:36] where the methods were placed in one paper with the formal evaluation appearing in a companion paper. The papers grouped as field trials of systems in routine use [1:35, 2:102, 307] might be better characterized as something other than research studies since their focus is upon methods.

This review of medical informatics research published in *JAMIA* indicates that the majority of effort is being devoted to projects involving the early stages of system development. Accordingly, most evaluation focuses upon need assessment and bench testing. This finding may represent sampling error in that validity and efficacy studies are likely to be accepted by less specialized biomedical journals. Nonetheless, we need a balance of research at each stage of development.[2] In particular, we need more controlled trials.[3]

The field of medical informatics will not gain widespread credibility until more innovations reach the level of maturity that permits researchers to document their efficacy through use in practice. As those studies become possible, a significant percentage should be published in an informatics journal, where they are most likely to reach people who need to be encouraged to do similar studies. The rest should go to very general biomedical journals, where they reach a large audience that needs to be aware of innovations that are ready for general use.

WILLIAM W. STEAD, MD

References ■

1. Stead WW, Haynes RB, Fuller S, et al. Designing medical informatics research and library resource projects to increase what is learned. *J Am Med Informatics Assoc.* 1994;1:28–33.
2. Friedman CP. Where's the science in medical informatics? *J Am Med Informatics Assoc.* 1995;2:65–7.
3. Tierney WM, Overhage JM, McDonald CJ. A plea for controlled trials in medical informatics. *J Am Med Informatics Assoc.* 1994;1:353–5.

Correspondence and reprints: William W. Stead, MD, The Annette and Irwin Eskind Biomedical Library, Vanderbilt University Medical Center, Nashville, TN 37232-8340. e-mail: bill.stead@mcmail.vanderbilt.edu

Received for publication: 9/29/95; accepted for publication: 9/29/95.

■