Evaluating Clinical Decision Support Systems

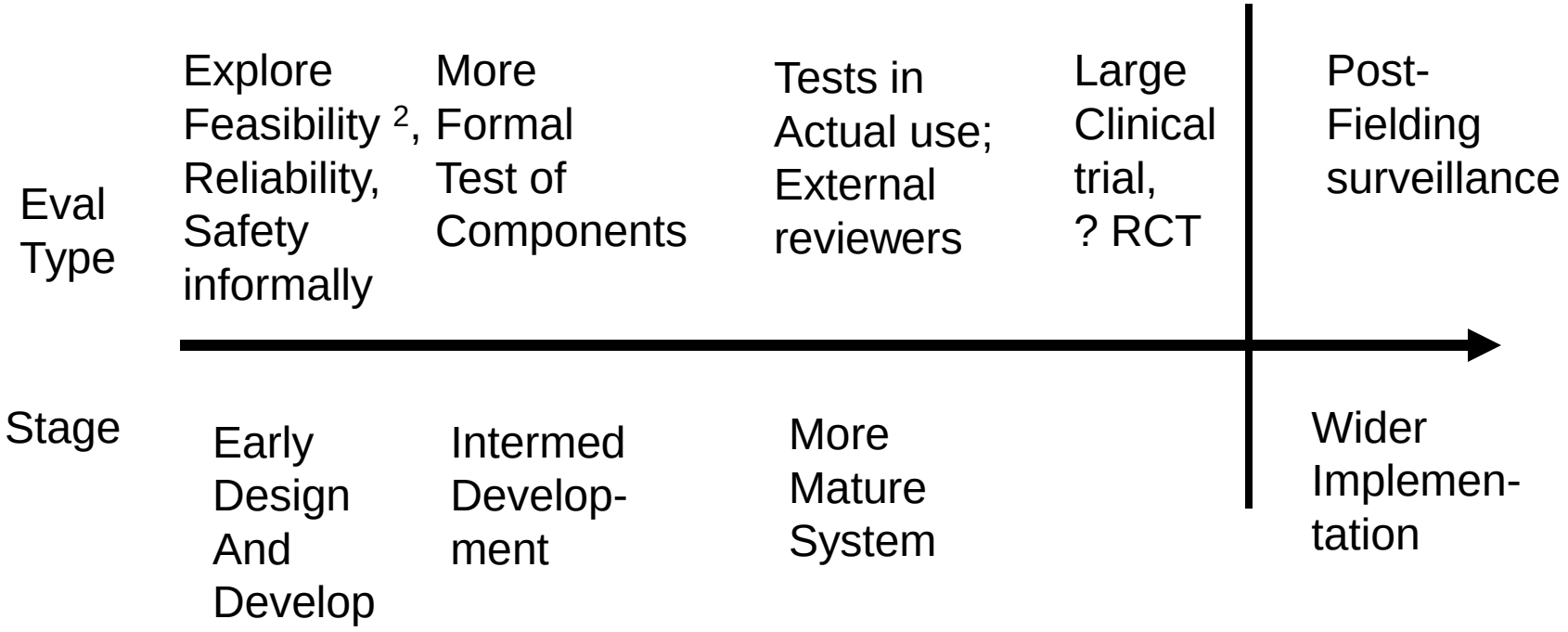
From Initial Design to Post-Deployment

Presented by Mary K. Goldstein, MD
VA Palo Alto Health Care System and Stanford University
VA HSR&D Cyber Seminar 12/16/08

Goals/Outline

- Lifecycle of development of clinical decision systems
- Evaluation methods appropriate to different stages of development
- A method for offline testing of accuracy of recommendations

Stages in Evaluating Clinical Decision Support Systems [1]



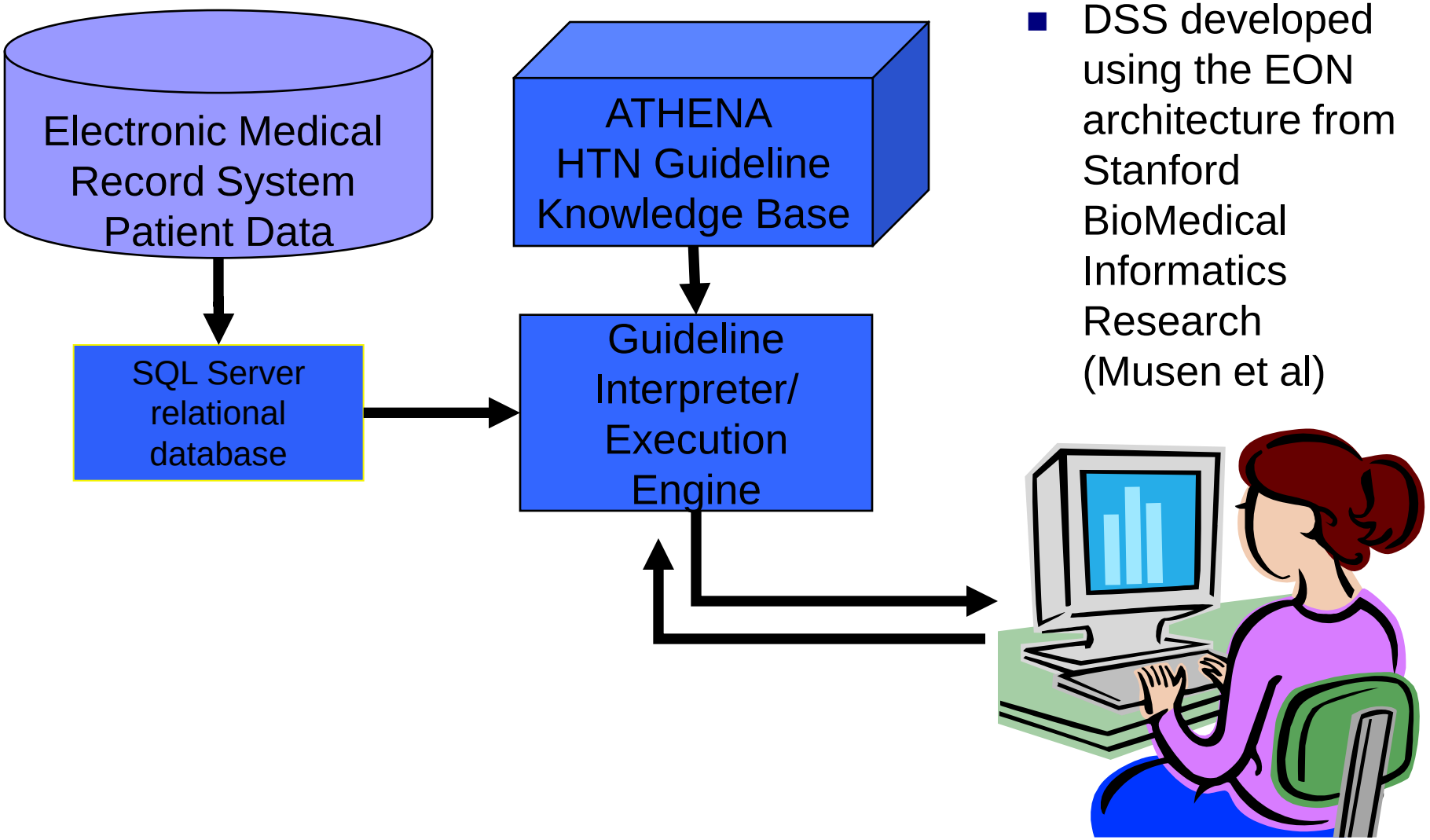
1. Figure developed largely from material in Miller RA JAMIA 1996
2. Use Cases

ATHENA Hypertension (HTN)

- **Clinical Domain:** Primary hypertension
  - JNC and VA Hypertension guidelines
- **Intended User:**
  - Primary care clinicians
- **Architecture:** EON Architecture for guideline-based information systems

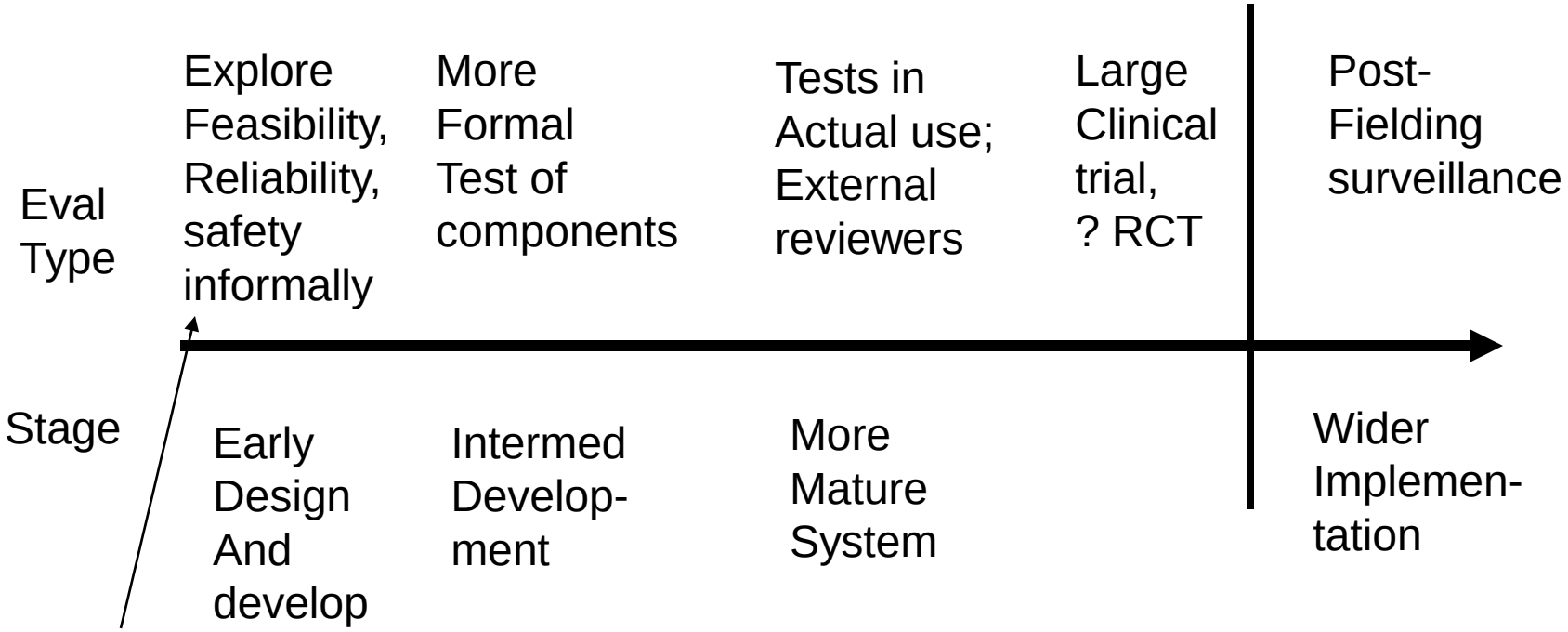
Goldstein MK, Coleman RW, Tu SW, et al. Translating research into practice. JAMIA 2004 Sep-Oct;11(5):368-76.

CDSS to Evaluate: ATHENA-HTN



- DSS developed using the EON architecture from Stanford BioMedical Informatics Research (Musen et al)

Stages in Evaluating Clinical Decision Support Systems (CDSS)



Goldstein, M.K., et al., *Patient Safety in Guideline-Based Decision Support for Hypertension Management: ATHENA DSS.* JAMIA, 2002. 9(6 Suppl): S11-6.

Testing Health IT for Patient Safety

- *"Latent errors or system failures pose the greatest threat to safety in a complex system because they lead to operator errors."*

  - Kohn LT, Corrigan JM, Donaldson MS, editors. To Err is Human: Building a safer health system. Washington, D.C.: National Academy Press; 2000.

Patient Safety in New Health IT

- New computer systems have potential to reduce errors…
  - But also potential to create new opportunities for error

Errors due to new Health IT

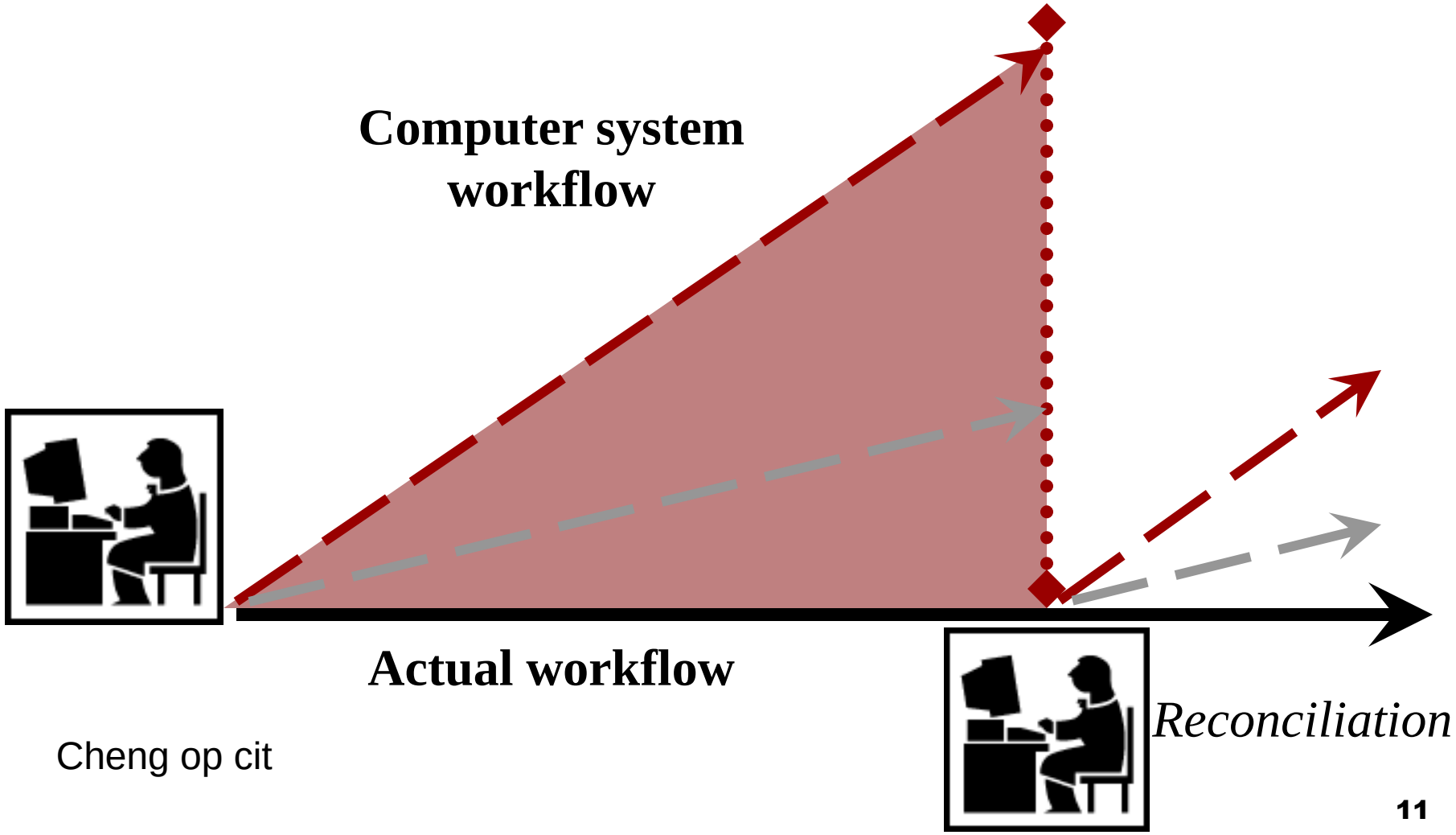
- Studies of accidents have shown that new computer systems can affect human problem solving in ways that contribute to errors
  - data overload
    - computer collects and displays information out of proportion to human ability to use it effectively
  - "automation surprises"
    - bar code administration unobservable action
      - Woods DD, Patterson ES et al. Can we ever escape from data overload? Human Factors & Ergonomics Soc 43rd Annual Meeting 1999.
      - Sarter NB, Woods DD. Hum Factors 2000.
      - Goldstein, M.K., et al., *Patient safety in guideline-based decision support for hypertension management: ATHENA DSS.* J Am Med Inform Assoc, 2002. **9**(6 Suppl): p. S11-6 (summarizes)

Computerized Physician Order-Entry (CPOE) in an Intensive Care Unit (ICU)

- Qualitative evaluation of introduction of mandatory CPOE to an ICU (next 2 slides)

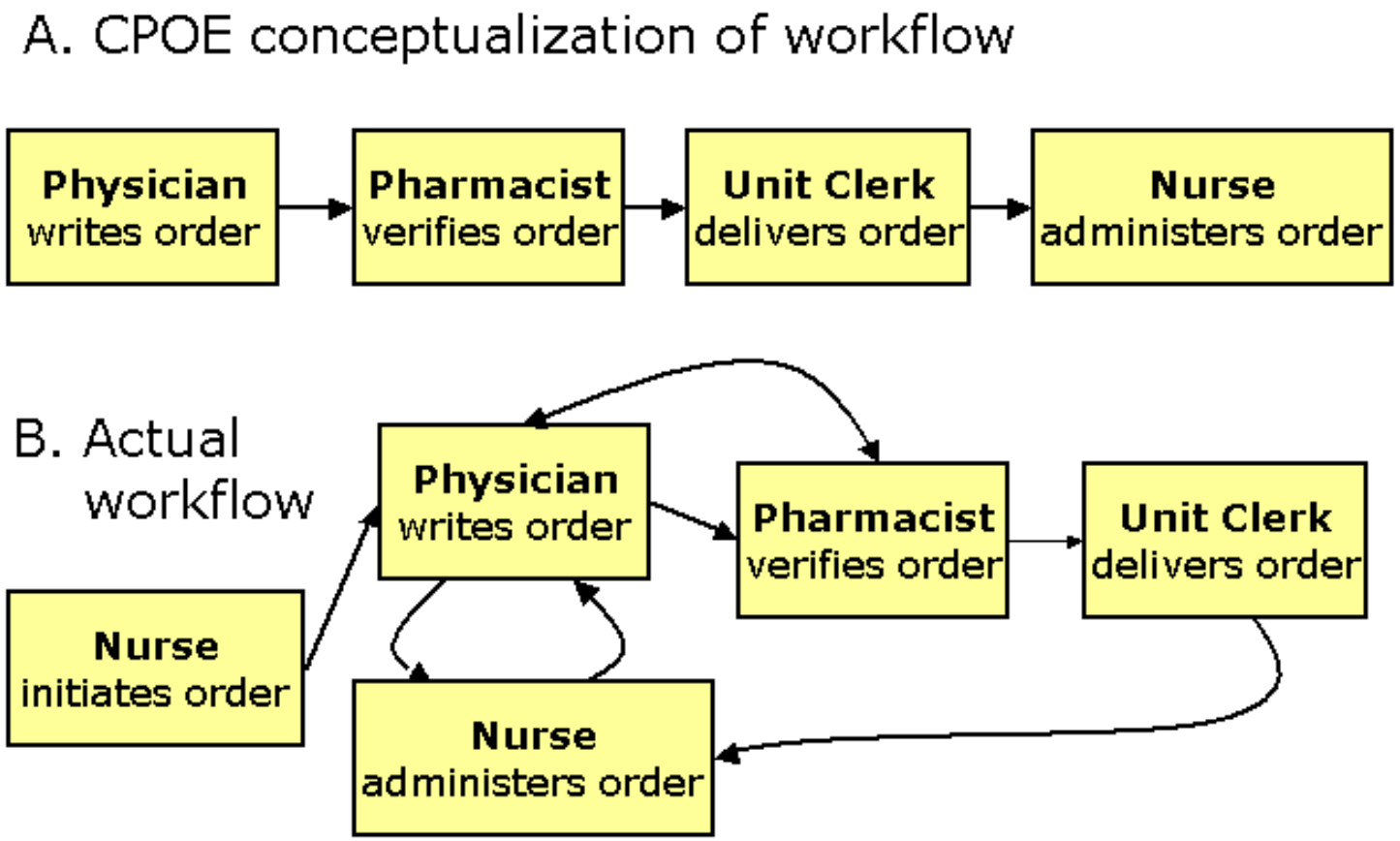
Cheng, C.H., et al., *The Effects of CPOE on ICU Workflow: An Observational Study.* Proc AMIA Symp, 2003: p. 150-4.

Computer system workflow diverges from actual workflow



Cheng op cit

Coordination redundancy: (Cheng op cit) Entering and interpreting orders



In 97 interruptions of RN to MD, 25% were reminders

Importance of Iterative Design

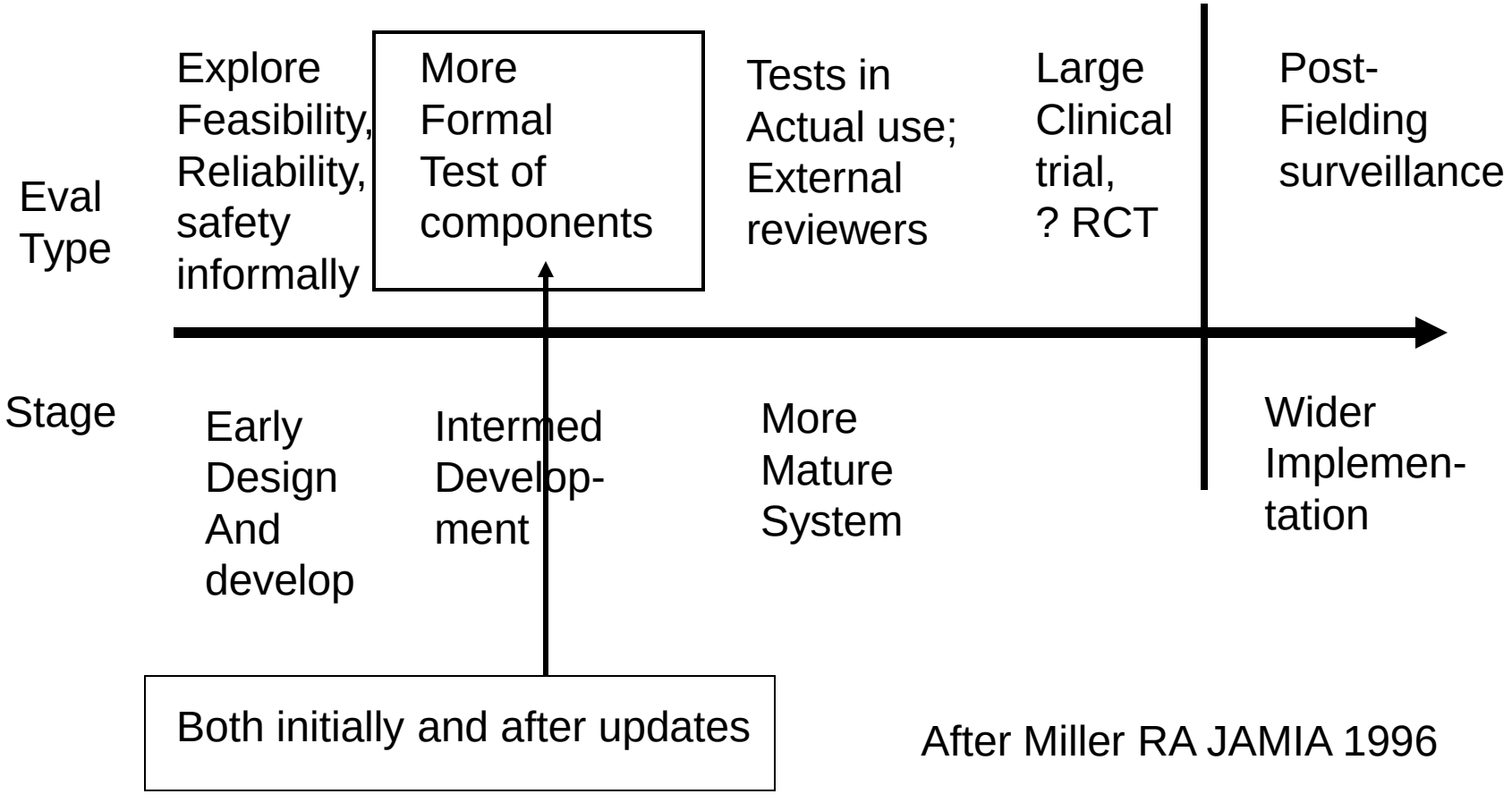
- Findings such as above from accident reports suggest the need for thorough testing of new information technology
  - accuracy, and also
  - usability, usefulness, understanding
- Project budgets and timelines should be constructed to allow for redesign and retesting after initial testing
  - Iterative design/testing cycles

Safety Testing Clinical Decision Support Systems

- "Before disseminating any biomedical information resource…designed to influence real-world practice decisions…check that it is safe…"
  - Drug testing in vitro before in vivo
- Information resource safety testing:
  - how often it furnishes incorrect advice

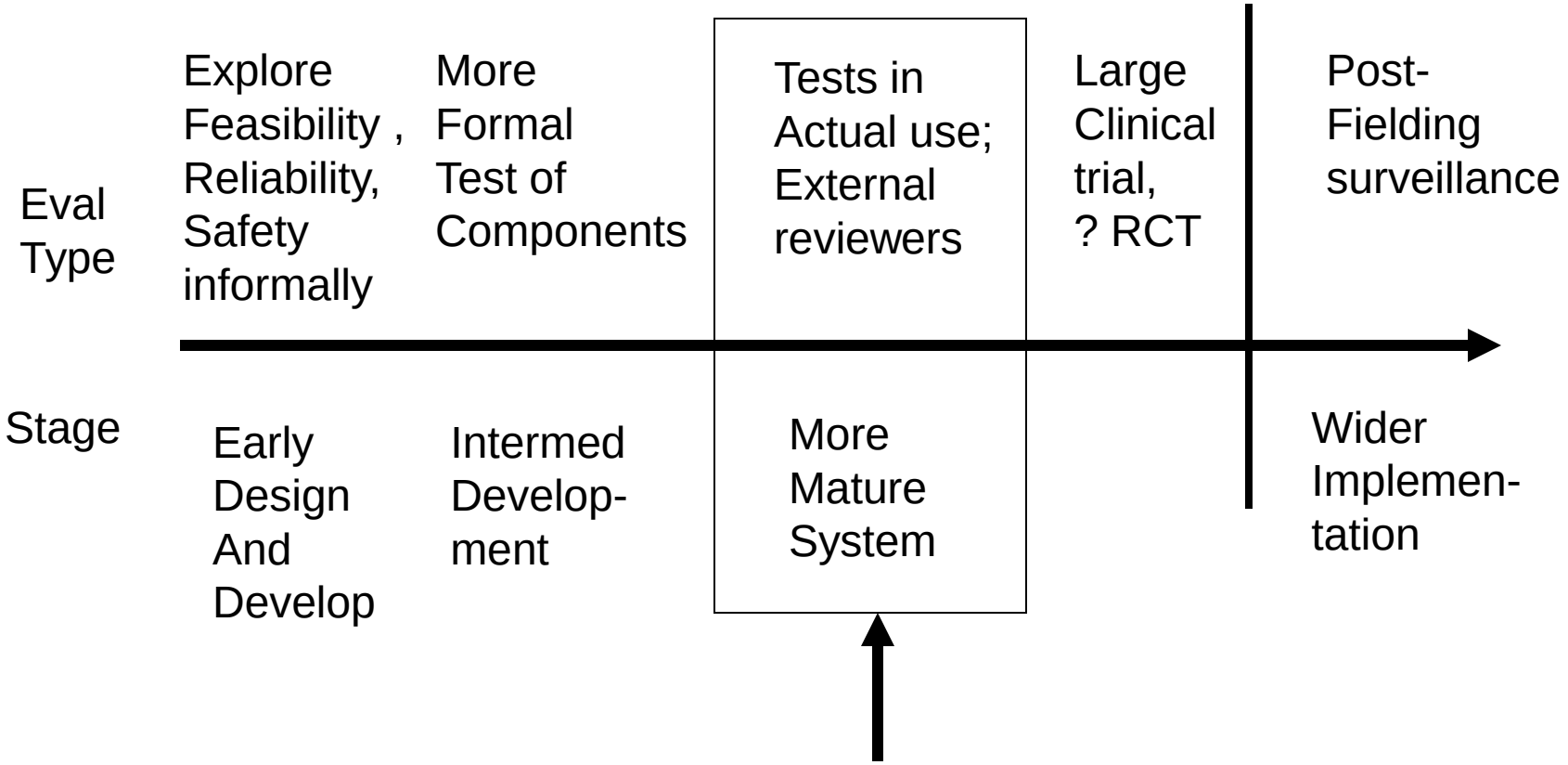
Friedman and Wyatt *Evaluation Methods in Biomedical Informatics* 2006

Stages in Evaluating Clinical Decision Support Systems



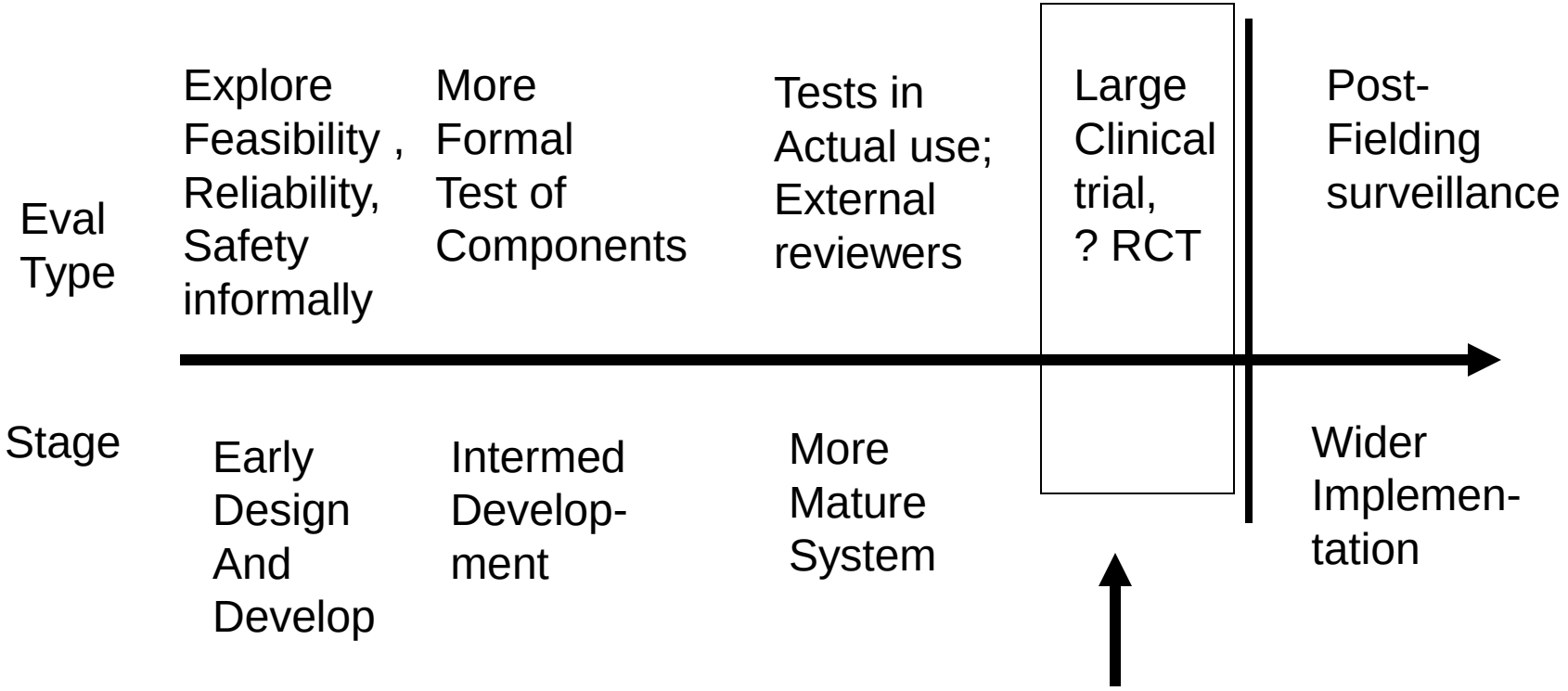
After Miller RA JAMIA 1996

Stages in Evaluating Clinical Decision Support Systems

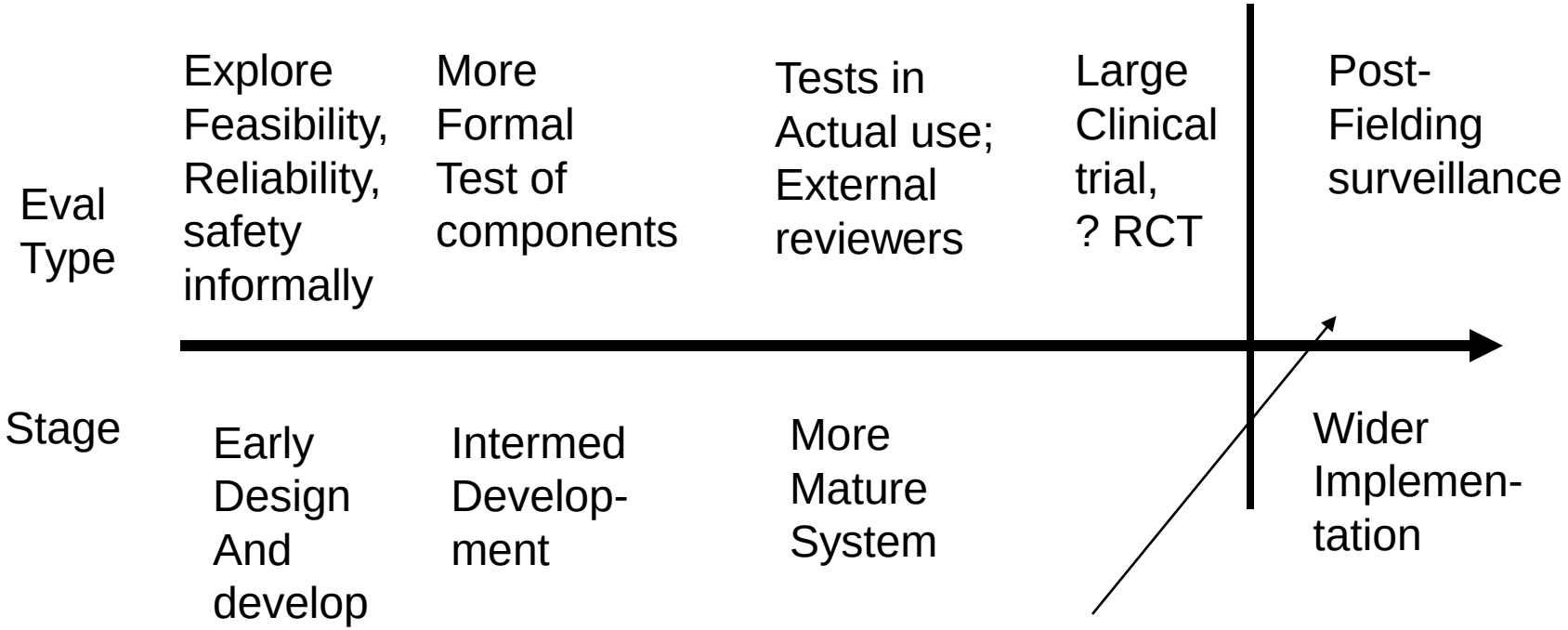


JAMIA 2004 op cit

Stages in Evaluating Clinical Decision Support Systems

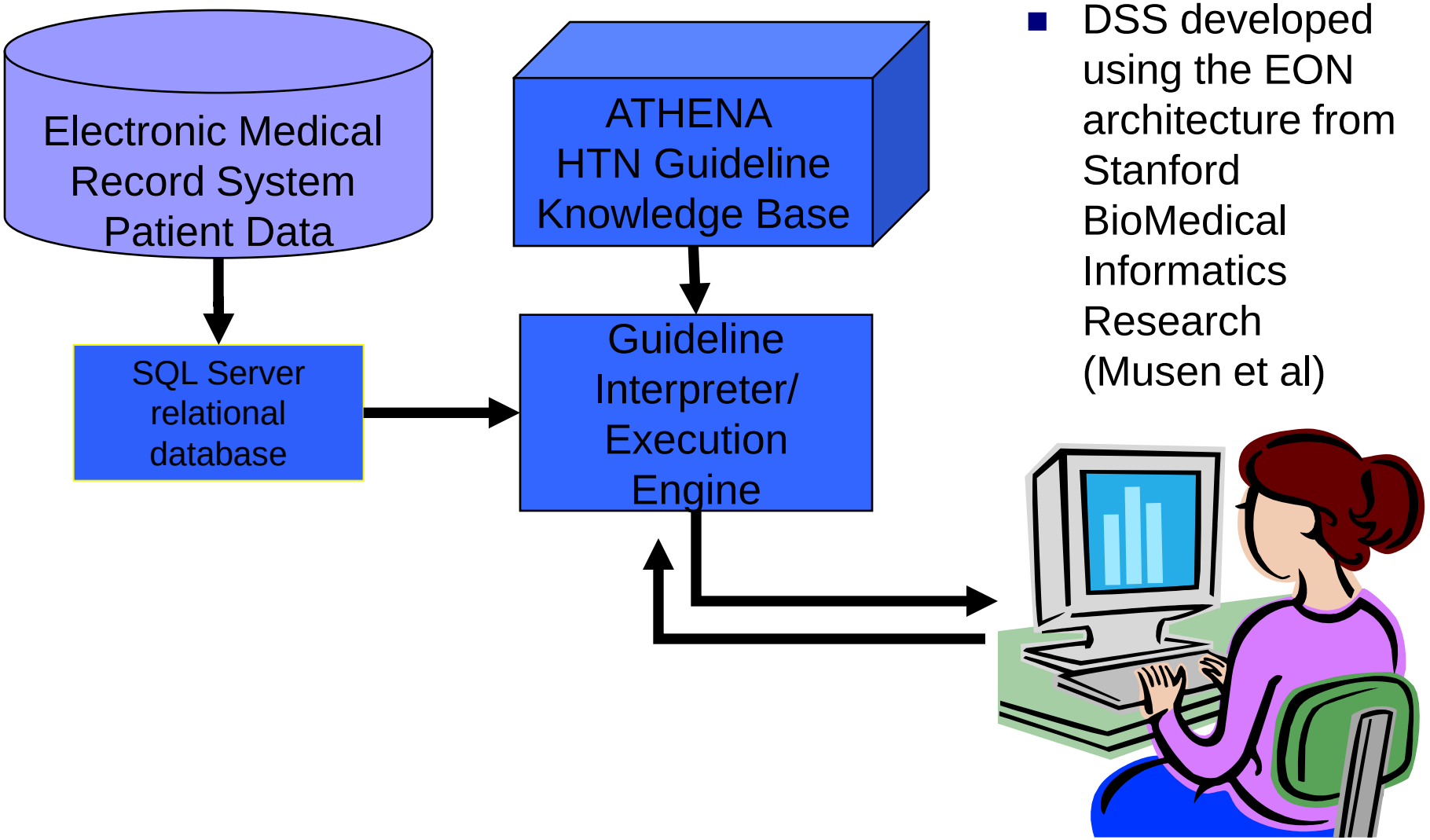


Stages in Evaluating Clinical Decision Support Systems (CDSS)



Chan AS et al Post Fielding Surveillance...
Advances in Patient Safety: From Research to Implementation. **Vol. 1. Research Findings**
AHRQ Publication Number 05-0021-1

CDSS to Evaluate: ATHENA-HTN



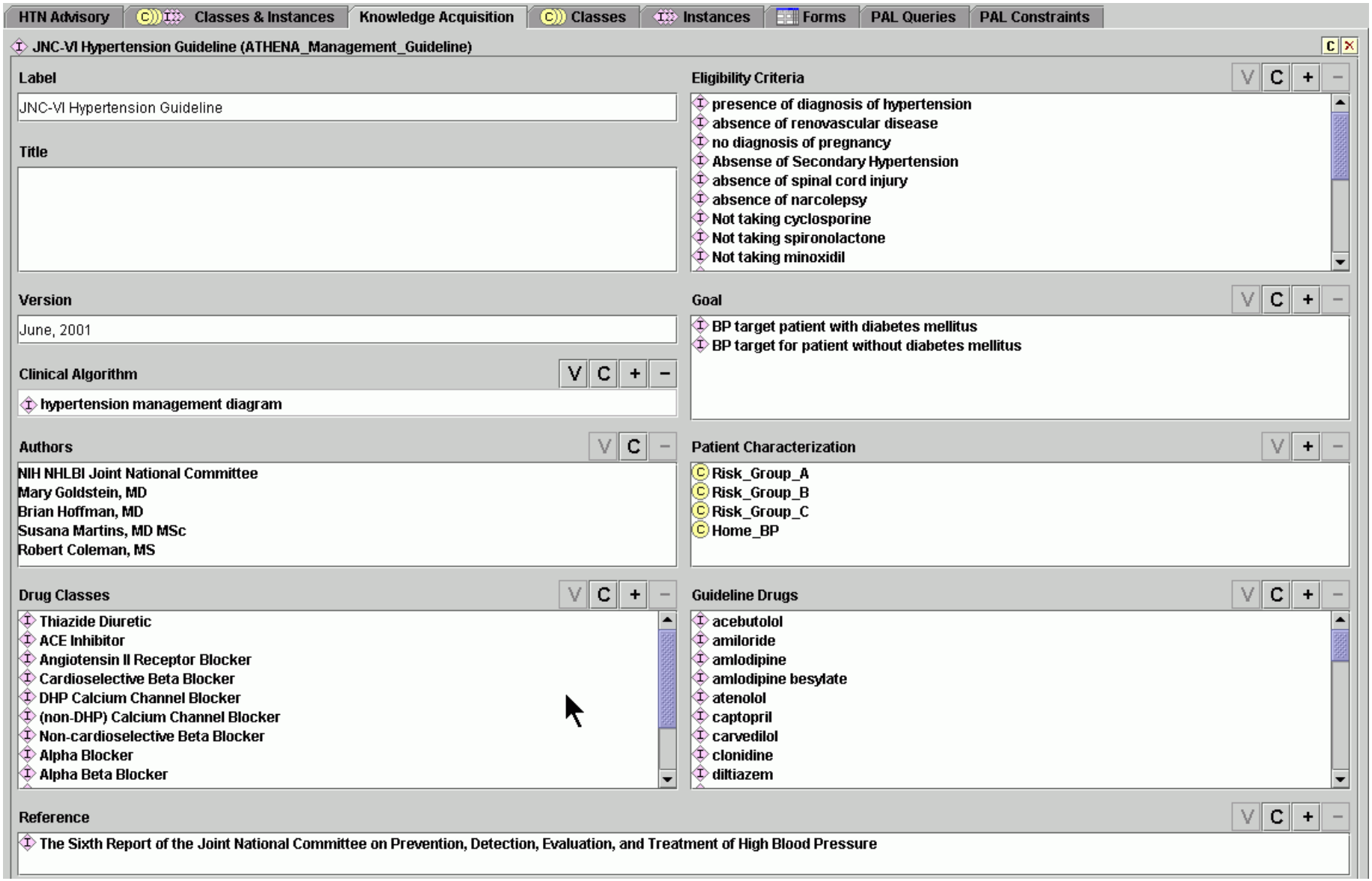
- DSS developed using the EON architecture from Stanford BioMedical Informatics Research (Musen et al)

Knowledge Base

- Protégé: ontology editor
  - Open source (http://protege.stanford.edu/)
- EON model for practice guidelines
- Focus for evaluation:
  - Eligibility criteria for including patients
  - Drug reasoning for drug recommendations

Tu SW, Musen MA. A Flexible Approach to Guideline Modeling. Proc AMIA Symp; 1999. 420-424

HTN Knowledge Base in Protégé



Guideline Execution Engine

- Applies the guideline as encoded in the knowledge base to the patient's data
- Generates set of recommendations

Tu SW, Musen MA. Proc AMIA Symp; 2000. 863-867

“The Art of Software Testing”

- False definition of testing
  - E.g., “Testing is the process of demonstrating that errors are not present”
- Testing should add value to the program
  - improve the quality
- Start with assumption program contains errors
  - A valid assumption for almost any program
- “Testing is the process of executing a program with the intent of finding errors.”

Myers G, Sandler C, Badgett T, Thomas T. The Art of Software Testing. 2nd Ed. John Wiley & Sons; 2004

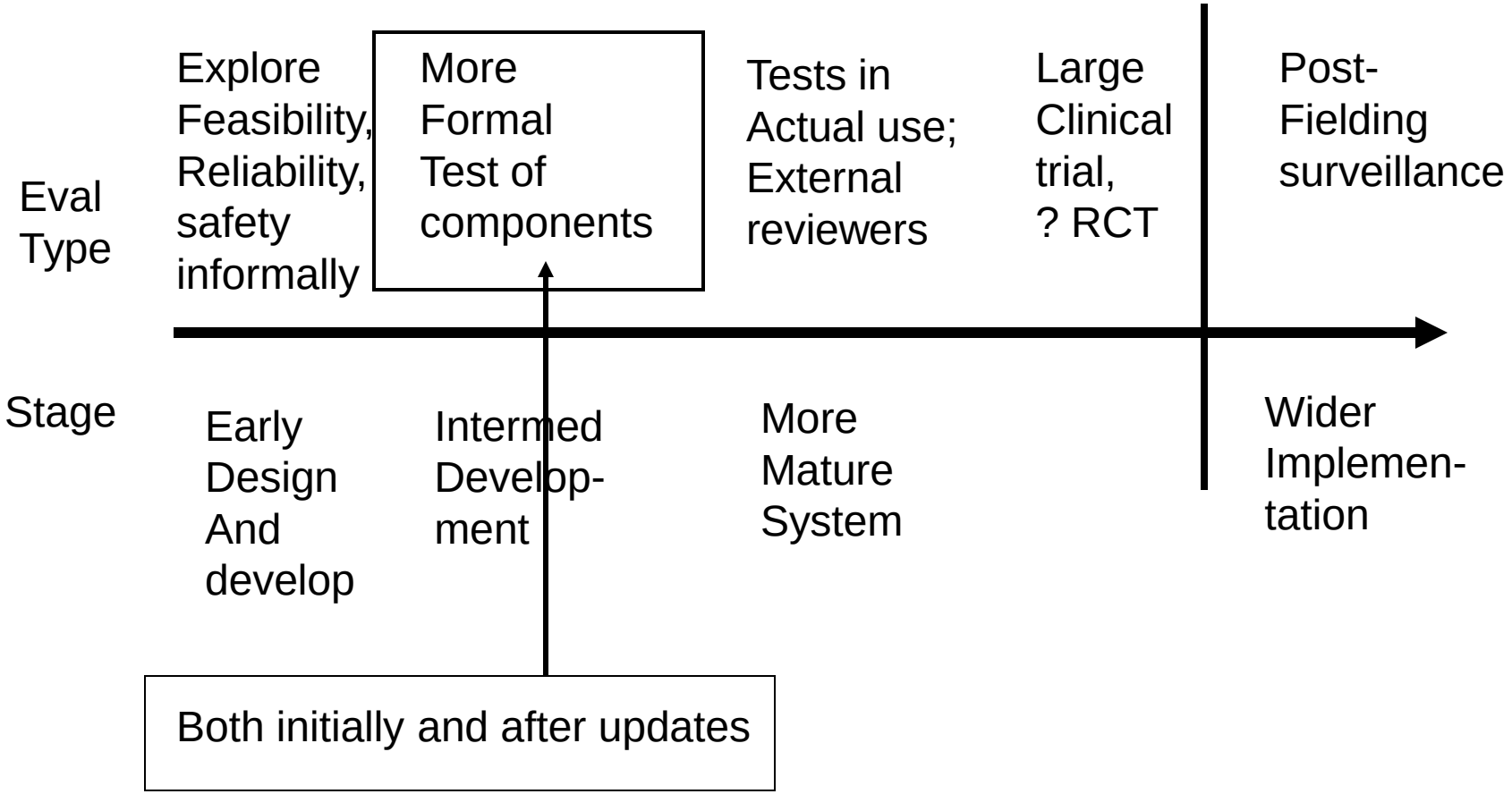
Software “Regression Testing” *

- Software updates and changes are particularly error-prone
- Changes may introduce errors into a previously well-functioning system
  - “regress” the system
- Desirable to develop a set of test cases with known correct output to run in updated systems before deployment

(* not statistical regression)

Myers et al op cit

Stages in Evaluating Clinical Decision Support Systems

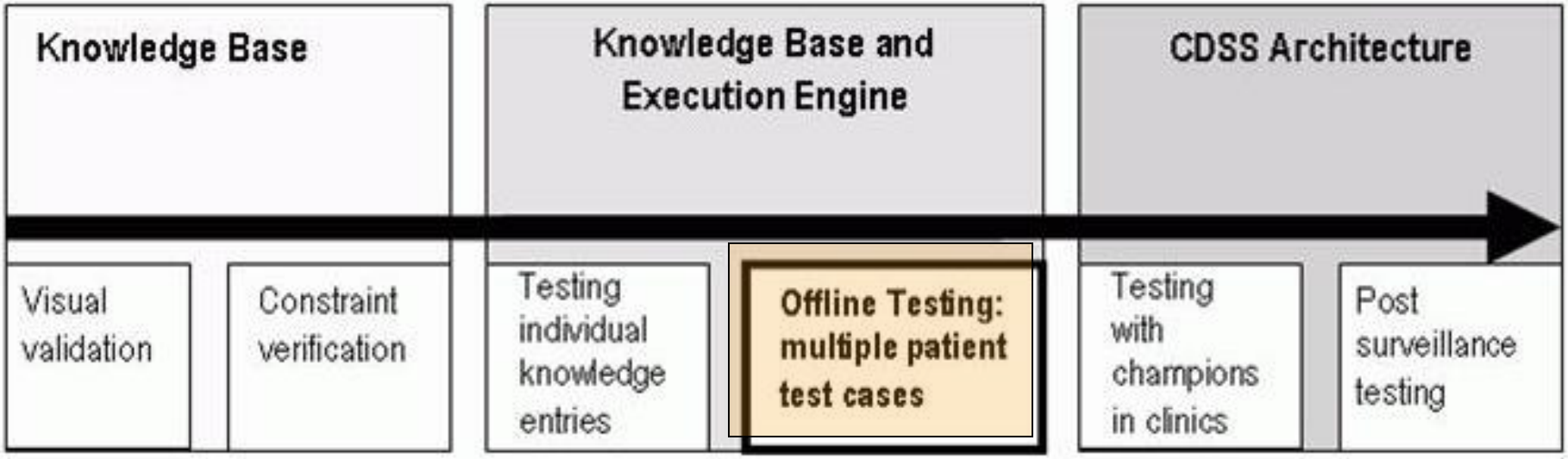


Our Testing at this Phase

The following slides are based on study reported in:

Martins, S.B., S. Lai, S.W. Tu, R. Shankar, S.N. Hastings, B.B. Hoffman, N. Dipilla, and M.K. Goldstein, *Offline Testing of the ATHENA Hypertension Decision Support System Knowledge Base to Improve the Accuracy of Recommendations.*
AMIA Annu Symp Proc, 2006: 539-43.

Clinical Decision Support System Accuracy Testing Phases



Further breakdown of steps as they apply to testing systems built on knowledge bases. Lin N op cit focuses on the highlighted phase of testing.

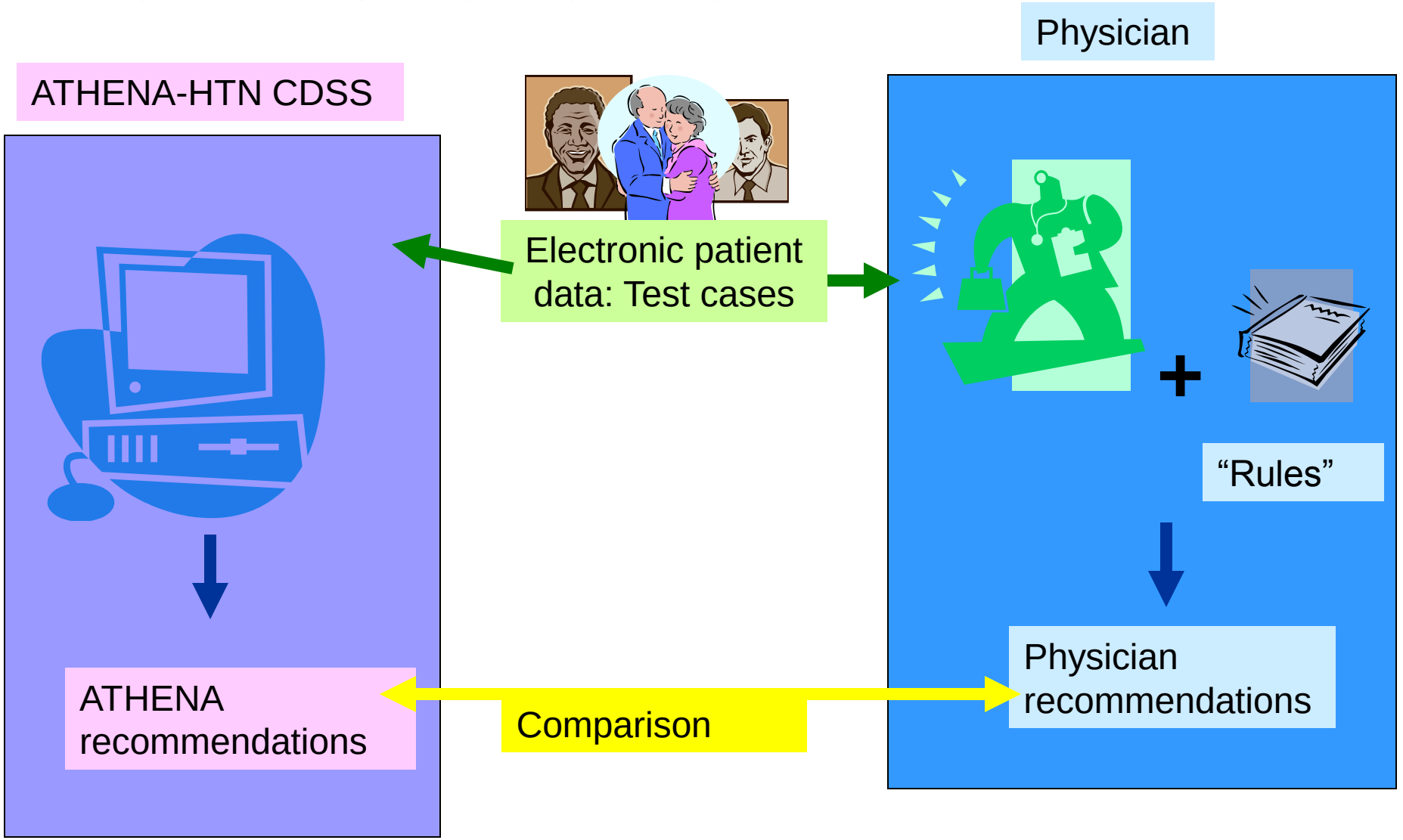
Objectives for Offline Testing of Accuracy of Recommendations

- Test the **knowledge base and the execution engine** after an update to the knowledge base and prior to clinical deployment of the updated system
  - to detect errors and improve quality of system
- Establish correct output (answers) for set of test cases

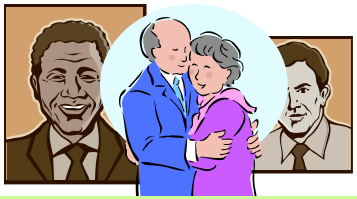
Comparison Method

- Comparing ATHENA vs MD output:
  - Automated comparison for discrepancies
  - Manual review of all cases
- Reviewing discrepancies
  - Meeting with physician evaluator
  - Adjudication by third party when categorizing discrepancies

Methods: Overview



Selection of Test Cases



Electronic patient data

100 cases from real patient data, 20 cases for each category:

- Heart failure
- Diabetes
- Diabetes & heart failure
- Coronary artery disease
- Uncomplicated hypertension

“Rules” Document

“Rules”

- Description of encoded guideline knowledge in narrative form
  - Resolving ambiguities in guideline (Tierney et al)
  - Defining scope of knowledge (boundaries of program)

Example of a boundary specification:

Heart failure: Although diuretics are used as antihypertensive agents, the management of diuretics in heart failure is primarily for volume management and is beyond the scope of this hypertension program.

Physician Evaluator (MD)

- Internist with experience in treating hypertension in primary care setting
- No previous involvement with ATHENA project
- Studied “Rules” and clarified any issues
- Had “Rules” and original guidelines available during evaluation of test cases

Elements examined

- **Patient eligibility**
  - Did patient meet ATHENA exclusion criteria?
- **Drug recommendations**
  - List of all possible anti-hypertensive drug recommendations concordant with guidelines
    - Drug dosage increases
    - Addition of new drugs
    - Drug substitutions
- **Comments by MD**

Comparison Method

- Comparing ATHENA vs MD ouput:
  - Automated comparison for discrepancies
  - Manual review of all cases
- Reviewing discrepancies
  - Meeting with physician evaluator
  - Adjudication by third party when categorizing discrepancies

Results: Drug Recommendations

- 92 eligible test cases:

	Drug recommendations			
	N	Mean	Median	Range
ATHENA	181	2	2	0-5
MD	184	2	2	0-5

- 27 discrepant drug recommendations
  - 8 due to problems with MD interpretation of pharmacy text (SIG in terms understood by pharmacists not MDs)
  - 19 other discrepancies:
    - ATHENA more comprehensive in recommendations (eg MD stopped after identifying some rec's w/o listing all) (15)
    - Ambiguity in the Rules being interpreted by MD (3)
    - Rules document contained a rec not encoded in KB (1)

MD Comments: 10

- 3 comments identified new boundary
  - E.g., BB Sotalol as anti-arrhythmic drug
- 7 comments identified known boundaries not explicit in Rules document
  - Drug dose decrease
  - Check for prescribed drugs that cause hypertension
  - Managing potassium supplement doses

Successful Test

- A successful test is one that finds errors
  - so that you can fix them

Myers et al, op cit

ATHENA Knowledge Base: Updates

- 3 updates made:
  - Added new exclusion criteria
  - Hydrochlorothiazide was added as a relative indication for patients on multi-drug regimen
  - Sotalol was re-categorized as an anti-arrhythmic drug

Set of “Gold Standard” Test Cases

- Iteration between clinician review and system output
- Same test cases for bug fixes and elaborations in areas that don’t affect the answers to test cases
- Change gold standard answers to test cases when the GL changes
  - i.e., when what you previously thought was correct is no longer correct (the clinical trial evidence and guidelines change over time)

Important features of Offline Testing Method

- Challenging CDSS with real patient data
- Clinician not involved in project: “fresh view”

Additional observation

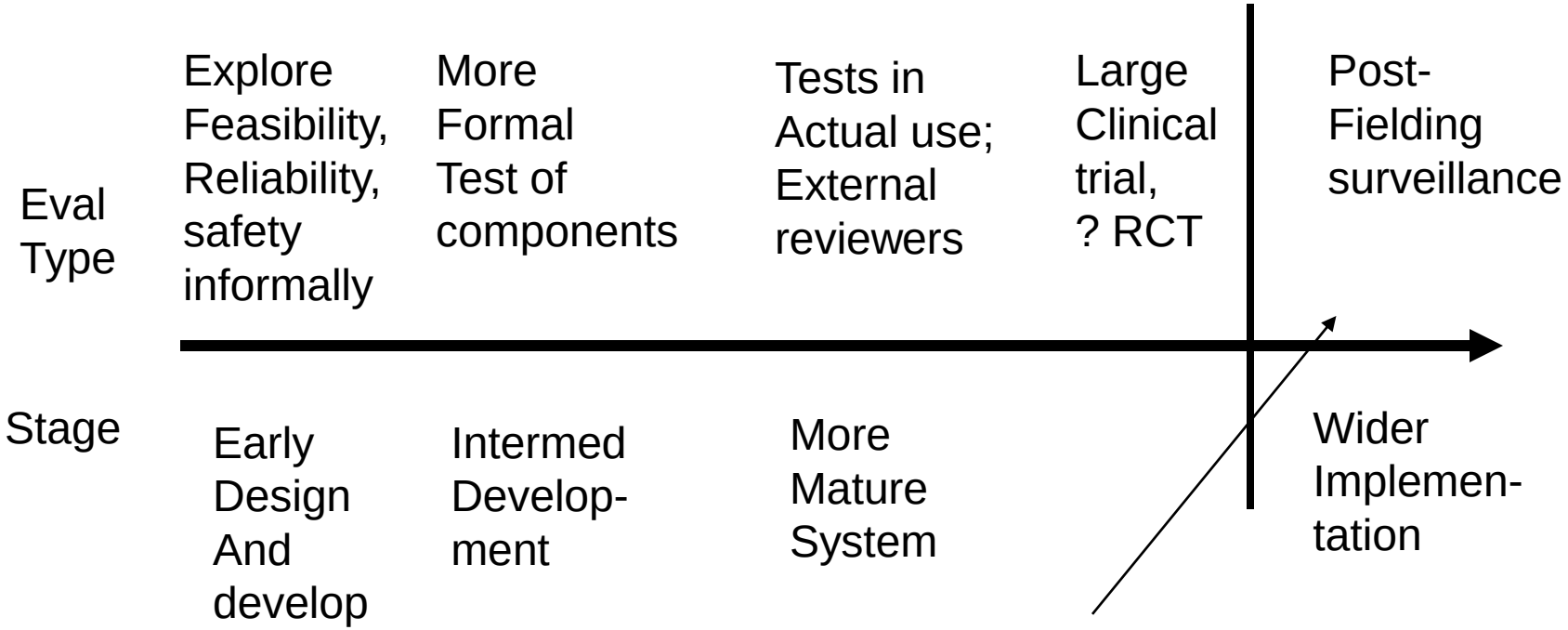
- Difficulty of maintaining a separate “Rules” document that describes encoded knowledge

Benefits of the Offline Testing

- Offline testing method was successful in identifying “errors” in ATHENA’s Knowledge base
- Program boundaries were better defined
- Updates made improving accuracy before deployment
- Gold standard answers to test cases

Offline Testing of the ATHENA Hypertension Decision Support System Knowledge Base to Improve the Accuracy of Recommendations.
Martins SB, Lai S, Tu SW, Shankar R, Hastings SN, Hoffman BB, Dipilla N, Goldstein MK.
AMIA Annu Symp Proc. 2006;539-43

Reminder to continue monitoring after deployment



Chan AS et al Post Fielding Surveillance...
Advances in Patient Safety: From Research to Implementation. **Vol. 1. Research Findings**
AHRQ Publication Number 05-0021-1

Books on Evaluation

- For software testing:
  - The Art of Software Testing. Eds Myers GJ et al. Wiley and Sons. 2004 (2nd edition)
- For everything else about evaluation of health informatics technologies
  - Evaluation Methods in Biomedical Informatics. Friedman CP and Wyatt JC. Springer 2006 (2nd edition)

STARE-HI Principles

Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI)

A comprehensive list of principles relevant for properly describing Health Informatics evaluations in publications

- endorsed by
  - European Federation of Medical Informatics (EFMI) council
  - American Medical Informatics Association (AMIA) Working Group (WG) on Evaluation
- Watch for further information on STARE-HI

VA PAHCS
Veterans Affairs Palo Alto Health Care System

Stanford University
School of Medicine