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

<table>
<thead>
<tr>
<th>Eval Type</th>
<th>Stage</th>
<th>Early Design And Develop</th>
<th>Intermed Development</th>
<th>More Mature System</th>
<th>Tests in Actual use; External reviewers</th>
<th>Large Clinical trial, ? RCT</th>
<th>Post-Fielding surveillance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explore Feasibility</td>
<td>Early Design And Develop</td>
<td>Early Design And Develop</td>
<td>Intermed Development</td>
<td>More Mature System</td>
<td>Tests in Actual use; External reviewers</td>
<td>Large Clinical trial, ? RCT</td>
<td>Post-Fielding surveillance</td>
</tr>
<tr>
<td>Reliability, Safety</td>
<td></td>
<td>Early Design And Develop</td>
<td>Intermed Development</td>
<td>More Mature System</td>
<td>Tests in Actual use; External reviewers</td>
<td>Large Clinical trial, ? RCT</td>
<td>Post-Fielding surveillance</td>
</tr>
<tr>
<td>informally</td>
<td></td>
<td>Early Design And Develop</td>
<td>Intermed Development</td>
<td>More Mature System</td>
<td>Tests in Actual use; External reviewers</td>
<td>Large Clinical trial, ? RCT</td>
<td>Post-Fielding surveillance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Early Design And Develop</td>
<td>Intermed Development</td>
<td>More Mature System</td>
<td>Tests in Actual use; External reviewers</td>
<td>Large Clinical trial, ? RCT</td>
<td>Post-Fielding surveillance</td>
</tr>
</tbody>
</table>

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

CDSS to Evaluate: ATHENA-HTN

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

Flowchart:
- Electronic Medical Record System Patient Data
- SQL Server relational database
- ATHENA HTN Guideline Knowledge Base
- Guideline Interpreter/Execution Engine
- Computer user interface
## Stages in Evaluating Clinical Decision Support Systems (CDSS)

<table>
<thead>
<tr>
<th>Eval Type</th>
<th>Stage</th>
<th>Eval</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Design And develop</td>
<td>Early</td>
<td>Explore</td>
<td>Feasibility, Reliability, safety informally</td>
</tr>
<tr>
<td>Intermed Development</td>
<td>Intermed</td>
<td>More</td>
<td>Formal Test of components</td>
</tr>
<tr>
<td>More Mature System</td>
<td>More</td>
<td>Tests in Actual use; External reviewers</td>
<td></td>
</tr>
<tr>
<td>Large Clinical trial, ? RCT</td>
<td>More</td>
<td>Post-Fielding surveillance</td>
<td></td>
</tr>
<tr>
<td>Wider Implementation</td>
<td>Wider</td>
<td>Implmen-tation</td>
<td></td>
</tr>
</tbody>
</table>

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.”

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

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

Computer system workflow diverges from actual workflow

Computer system workflow

Actual workflow

Reconciliation

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

A. CPOE conceptualization of workflow

- Physician writes order
- Pharmacist verifies order
- Unit Clerk delivers order
- Nurse administers order

B. Actual workflow

- Nurse initiates order
- Physician writes order
- Pharmacist verifies order
- Nurse administers order
- Unit Clerk delivers order

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

<table>
<thead>
<tr>
<th>Stage</th>
<th>Eval Type</th>
<th>Type</th>
<th>Eval</th>
<th>Test</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Early Design And develop</td>
<td>Explore Feasibility, Reliability, safety informally</td>
<td>More Formal Test of components</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intermed Development</td>
<td>Tests in Actual use; External reviewers</td>
<td>More Mature System</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>More Mature System</td>
<td>Large Clinical trial, ? RCT</td>
<td>Post-Fielding surveillance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Both initially and after updates</td>
<td>After Miller RA JAMIA 1996</td>
<td>Wider Implementation</td>
</tr>
</tbody>
</table>
Stages in Evaluating Clinical Decision Support Systems

<table>
<thead>
<tr>
<th>Stage</th>
<th>Eval Type</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>Early Design And Develop</td>
<td>Explore Feasibility, Reliability, Safety informally</td>
<td>Tests in Actual use; External reviewers</td>
</tr>
<tr>
<td>Intermed</td>
<td>Intermed Development</td>
<td>More Formal Test of Components</td>
<td>Large Clinical trial, ? RCT</td>
</tr>
<tr>
<td>More Mature</td>
<td></td>
<td></td>
<td>Post-Fielding surveillance</td>
</tr>
<tr>
<td>Wider</td>
<td></td>
<td></td>
<td>Wider Implementation</td>
</tr>
</tbody>
</table>

JAMIA 2004 op cit
Stages in Evaluating Clinical Decision Support Systems

<table>
<thead>
<tr>
<th>Stage</th>
<th>Early Design And Develop</th>
<th>Intermed Development</th>
<th>More Mature System</th>
<th>Tests in Actual use; External reviewers</th>
<th>More Formal Test of Components</th>
<th>Explore Feasibility, Reliability, Safety informally</th>
<th>Eval Type</th>
<th>Post-Fielding surveillance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Clinical trial, ? RCT</td>
<td>Wider Implementation</td>
<td>Wider Implementation</td>
<td>Wider Implementation</td>
<td>Wider Implementation</td>
<td>Wider Implementation</td>
<td>Wider Implementation</td>
<td>Wider Implementation</td>
<td>Wider Implementation</td>
</tr>
</tbody>
</table>
Stages in Evaluating Clinical Decision Support Systems (CDSS)

<table>
<thead>
<tr>
<th>Eval Type</th>
<th>Early Design And develop</th>
<th>Intermed Development</th>
<th>More Mature System</th>
<th>Tests in Actual use; External reviewers</th>
<th>Large Clinical trial, ? RCT</th>
<th>Post-Fielding surveillance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage</td>
<td>Early Design And develop</td>
<td>Intermed Development</td>
<td>More Mature System</td>
<td>Tests in Actual use; External reviewers</td>
<td>Large Clinical trial, ? RCT</td>
<td>Post-Fielding surveillance</td>
</tr>
</tbody>
</table>

Chan AS et al Post Fielding Surveillance...

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)

Diagram:
- Electronic Medical Record System
  - Patient Data
  - SQL Server relational database
- ATHENA HTN Guideline Knowledge Base
- Guideline Interpreter/Execution Engine

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

HTN Knowledge Base in Protégé

Label
JNC-VI Hypertension Guideline

Title

Eligibility Criteria
- presence of diagnosis of hypertension
- absence of renovascular disease
- no diagnosis of pregnancy
- Absence of Secondary Hypertension
- absence of spinal cord injury
- absence of narcolepsy
- Not taking cyclosporine
- Not taking spironolactone
- Not taking minoxidil

Version
June, 2001

Clinical Algorithm
hypertension management diagram

Goal
- BP target patient with diabetes mellitus
- BP target for patient without diabetes mellitus

Authors
NIH NHLBI Joint National Committee
Mary Goldstein, MD
Brian Hoffman, MD
Susana Martins, MD MSc
Robert Coleman, MS

Patient Characterization
- Risk_Group_A
- Risk_Group_B
- Risk_Group_C
- Home_BP

Drug Classes
- Thiazide Diuretic
- ACE Inhibitor
- Angiotensin II Receptor Blocker
- Cardiodefective Beta Blocker
- DHP Calcium Channel Blocker
- (non-DHP) Calcium Channel Blocker
- Non-cardiodefective Beta Blocker
- Alpha Blocker
- Alpha Beta Blocker

Guideline Drugs
- acetazolamide
- amiloride
- amiloridine
- amiloridine besylate
- atenolol
- captopril
- carvedilol
- clonidine
- diltiazem

Reference
The Sixth Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure
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.”

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

<table>
<thead>
<tr>
<th>Eval Type</th>
<th>Stage</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explore</td>
<td>Early Design And develop</td>
<td>Early</td>
<td>Explore feasibility, reliability, safety informally</td>
</tr>
<tr>
<td></td>
<td>Intermed Development</td>
<td>Intermed</td>
<td>Tests in component; more formal test of components.</td>
</tr>
<tr>
<td></td>
<td>More Mature System</td>
<td>More Maturity</td>
<td>Tests in actual use; external reviewers.</td>
</tr>
<tr>
<td></td>
<td>Post-Fielding surveillance</td>
<td>Wider Implementation</td>
<td>Large clinical trial, ? RCT.</td>
</tr>
</tbody>
</table>

Both initially and after updates
Our Testing at this Phase

The following slides are based on study reported in:

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

ATHENA-HTN CDSS

Electronic patient data: Test cases

Physician recommendations

Comparison

Physician

“Rules”
Selection of Test Cases

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

- Heart failure
- Diabetes
- Diabetes & heart failure
- Coronary artery disease
- Uncomplicated hypertension
“Rules” Document

- 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 output:
  - 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:

<table>
<thead>
<tr>
<th>Drug recommendations</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATHENA</td>
<td>181</td>
<td>2</td>
<td>2</td>
<td>0-5</td>
</tr>
<tr>
<td>MD</td>
<td>184</td>
<td>2</td>
<td>2</td>
<td>0-5</td>
</tr>
</tbody>
</table>

- 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. 
Reminder to continue monitoring after deployment

Stage
- Early Design
- Intermed Development
- More Mature System
- Large Clinical trial, ? RCT
- Wider Implementation

Eval Type
- Explore Feasibility, Reliability, safety informally
- More Formal Test of components
- Tests in Actual use; External reviewers
- Large Clinical trial, ? RCT
- Post-Fielding surveillance

Chan AS et al Post Fielding Surveillance...
Books on Evaluation

- For software testing:

- For everything else about evaluation of health informatics technologies
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