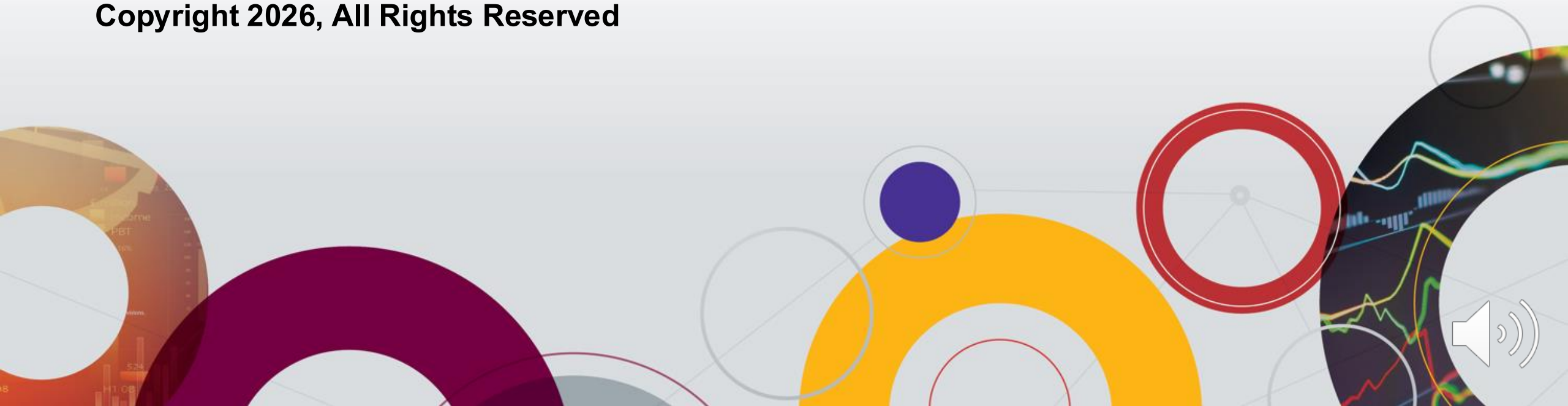


8.6a Evaluation and Future Directions

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Evaluation and future directions

Evaluation

- How valuable are systems to users?
- How well do systems and users perform?
- Applying IR techniques to electronic health records (EHRs)
- Generative IR

Future directions

Evaluation

Questions often asked

- Is system used?
- Are users satisfied?
- Do they find relevant information?
- Do they complete their desired task?

Most studied group is physicians, with systematic reviews of results (Hersh, 1998, Pluye, 2005; Maggio, 2019)

Much early work on usage and satisfaction reflects older systems before ubiquitous Internet, EHRs, mobile devices, etc.

Most IR evaluation research has focused on retrieval of relevant documents, which may not capture full spectrum of usage

- Often consists of challenge evaluations that develop “test collections” – best known is (non-medical) [Text Retrieval Conference \(TREC\)](#) (Voorhees, 2005)
 - [25th anniversary](#) featured many talks, including overview of [biomedically oriented tracks](#)

Do they find relevant information?

Most common approach to evaluation – based on retrieved documents judged for relevance to topic of search

Usually measured by relevance-based measures of recall and precision

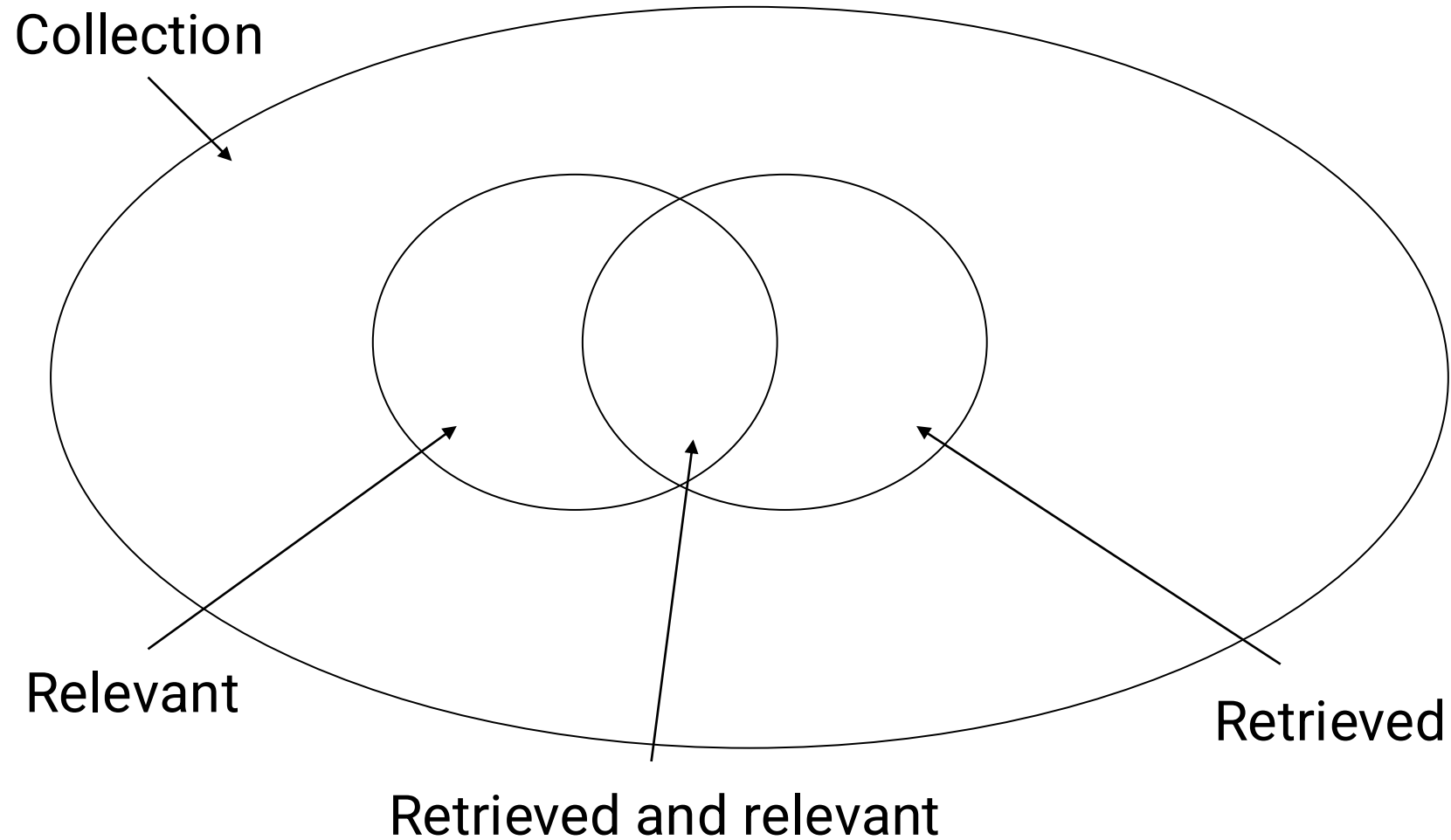
- Recall (R)

$$R = \frac{\text{\#retrieved and relevant documents}}{\text{\#relevant documents in collection}}$$

- Precision (P)

$$P = \frac{\text{\#retrieved and relevant documents}}{\text{\#retrieved documents}}$$

Visual depiction of recall and precision



Recall-precision example

Example:

- 100 known relevant documents
- 50 documents retrieved
- 25 documents retrieved are relevant

Recall = $25/100 = 25\%$

Precision = $25/50 = 50\%$

Comments about recall and precision

Usually do not know all relevant documents, so use relative recall of known relevant documents

There tends to be a trade-off between the two

“Relevance” can be an ambiguous notion (Hersh, 1994)

It is unclear whether they correlate with a user’s success in using an IR system

The proliferation of standard test collections leads to a great deal of research that excludes real users

How well do clinicians search? Early results from Haynes (1990)

Searcher Type	Recall	Precision
Novice clinicians	27%	38%
Expert clinicians	48%	48%
Librarians	49%	57%

Other findings

- Little overlap among retrieval sets
 - Searchers tended to find similar quantities of disparate relevant documents
- Novice searchers satisfied with results
 - Adequate information or ignorant bliss?

Extending evaluation beyond physicians and documents

Other clinicians

- Nurses – Rolye, 1995
- Pharmacists – Wanke, 1988
- Nurse practitioners – Hersh, 2000; Hersh, 2002
- Medical students – Hersh, 2002

Biomedical researchers

- Very little study of their use of IR systems
- Investigated by TREC Genomics Track (Hersh, 2006; Hersch, 2009)

Image retrieval – ImageCLEFmed (Hersh, 2006; Hersch, 2009)

Recall and precision studies yield useful results, but

Are searchers able to solve their information problems by using system?

- Some results research have used “task-oriented approach” to measure question-answering
- Hersh (2002) – use of MEDLINE to answer clinical questions
 - Medical students answered 34% of questions before system, 51% afterwards
 - Nurse practitioner students answered 34% of questions before system but did not change with system
 - Time to answer a question was ~30 minutes
 - No association of recall or precision with correct answering

Another task-oriented study

Westbrook (2005) – use of online evidence system

- Physicians answered 37% of questions before system, 50% afterwards
- Nurse specialists answered 18% of questions before system, 50% afterwards
- Those who had correct answers had higher confidence in their answers, but those not knowing answer initially had no difference in confidence whether answer right or wrong

Studies using task-oriented methodology (cont.)

Searching for answers in Wikipedia, UpToDate, and a digital textbook by medical students (Scaffidi, 2017)

- Wikipedia as good as UpToDate and better than digital textbook

Comparing question-answering with different retrieval systems and time limitations (van der Vegt, 2020)

- Better retrieval system (by non-user studies) did not lead to improved answering of questions
- Retrieval system improved answer correctness (33-35% → 53-54%)
- Time constraint (3 min vs. 6-9 min) increased number correct before searching but incorrect after

Task-oriented studies comparing modern systems

Comparing UpToDate (UTD) and Dynamed

Medical residents (Bradley-Ridout, 2021)

- Much previous experience and satisfaction with UTD
- Question correctness comparable – around 67%
- Time taken faster for UTD – 6 vs. almost 8 minutes
- More confidence in answers and satisfaction in use with UTD

Medical students, residents, and faculty (Baxter, 2022)

- Much previous experience and satisfaction with UTD
- Question correctness comparable but highly variable by question
- Time taken comparable for each – range 2.1-4.7 minutes
- More satisfaction with UTD

How do IR systems impact physician practice? (Pluye, 2004)

Qualitative study found four themes mentioned by physicians

- Recall – of forgotten knowledge
- Learning – new knowledge
- Confirmation – of existing knowledge
- Frustration – that system use not successful

Researchers also noted two additional themes

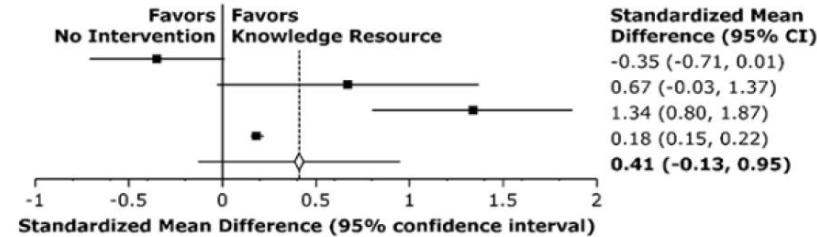
- Reassurance – that system is available
- Practice improvement – of patient-physician relationship

Impact on physician practice (Maggio, 2019)

Panel A: Knowledge/Skills Outcomes

Author, Year	Knowledge Resource	Outcome	Users (n)
Grad, 2005	InfoRet	Knowledge test	MD (33)
Bohicchio, 2006	Abx Guide ^a	Knowledge test	PG (12)
Skeate, 2007	Rep Sup ^a	Pathology test	PG, MS (28)
Reed, 2012	UTD, PIER	MOC exam	MD (15,148)

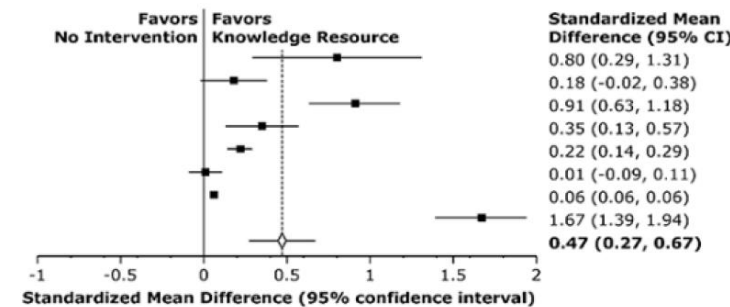
POOLED ($I^2=89\%$)



Panel B: Clinician Behavior Outcomes

Author, Year	Knowledge Resource	Outcome	Users (n)	Number of Patients
Greiver, 2005	Ang Soft ^a	Order test/consult	MD (17)	65
Ramnarayan, 2006 ^c	Isabel	Complication	PG (56)	104
Emery, 2007	GRAIDS ^a	Order test/consult	MD, NP (ns)	246
King, 2007 ^c	CEM ^a	Prescribe drug	PG, MS (ns)	316
Isaac, 2012	UTD	Safety/quality of care	Mixed (ns)	3322 ^b
Bonis, 2008	UTD	Safety/quality of care	Mixed (ns)	3515 ^b
Lyman, 2008	Epoc	Prescribe drug	MD (10,355)	13,957,797
Luther, 2016	SCAMP ^a	Order test/consult	MD (ns)	273

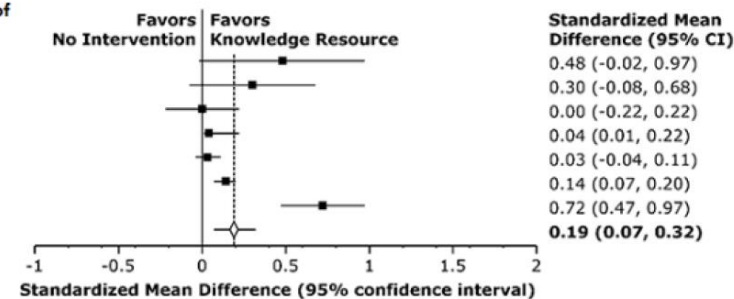
POOLED ($I^2=97\%$)



Panel C: Patient Effects Outcomes

Author, Year	Knowledge Resource	Outcome	Users (n)	Number of Patients
Greiver, 2005	Ang Soft ^a	Optimal management	MD (17)	65
Emery, 2007	GRAIDS ^a	Patient knowledge	MD, NP (ns)	127
King, 2007 ^c	CEM ^a	Length of stay	PG, MS (ns)	316
Bonis, 2008	UTD	Length of stay	Mixed (ns)	3515 ^b
Isaac, 2012	UTD	Length of stay	Mixed (ns)	3322 ^b
Kuhn, 2015	eAAP ^a	Patient symptoms	Mixed (ns)	3596
Luther, 2016	SCAMP ^a	Optimal management	MD (ns)	273

POOLED ($I^2=81\%$)



Consumer health searching: TREC Health Misinformation Track

Focus on consumer queries over Web content aiming to retrieve information that is correct and credible

Topics phrased as yes/no questions

Retrieved documents judged for

- Contain material that user might find useful in answering the question
- Contain information that supports/dissuades use of treatment
- Credibility

Tasks included

- Web retrieval – content that supports answers
- Answer prediction – predicting yes/no answer from question