ANNALS OF FAMILY MEDICINE

Online Supplementary Material

Alper BS, White DS, Ge B. Physicians answer more clinical questions and change clinical decisions more often with synthesized evidence: a randomized trial in primary care. *Ann Fam Med.* 2005;3:507-513.

http://www.annfammed.org/cgi/content/full/3/6/507/DC1

Supplemental Appendix. Level of Evidence Comparison

Brian S. Alper, MD, MSPH;[1] Richard O. Schamp, MD;[2] David S. White, MD;[3] Jennifer L. Hoock, MD[4]

[1]DynaMed and Department of Family and Community Medicine, University of Missouri-Columbia School of Medicine, Columbia, Mo
[2]Department of Community and Family Medicine, St Louis University School of Medicine, St Louis, Mo
[3]A private practice, Columbia, Mo
[4]Family Medicine Residency Network, University of Washington School of Medicine, St Louis, Mo

Introduction

As part of a Small Business Innovation Research (SBIR) award testing the feasibility of a clinical reference database (DynaMed), we had the following technical objectives:

1. Determine whether DynaMed increases the efficiency of answering clinical questions
2. Determine whether DynaMed contains valid answers

The first objective is described in the accompanying report, and the second objective is described in this online supplemental appendix.

We planned to have 3 physicians trained to interpret research methodology who did not have competing interests (ROS, DSW, JLH) compare the answers found in DynaMed with the answers found in a combination of 4 comparators to answer the question, "Does DynaMed provide answers with validity that meets or exceeds the validity of answers provided by key comparators?"

Methods

Our predefined goal was to determine whether the validity of answers found in DynaMed meets or exceeds that of the best answer found in 4 key comparators at least 50% of the time. For DynaMed answers to meet or exceed the others 60% of the time would require 387 questions for statistical significance at $\alpha = .01$ and $\beta = .05$.

We discovered complexities in applying our original protocol that necessitated protocol changes, which were approved by our institutional review board and National Science Foundation program officer. First, so many questions asked by primary care clinicians in the randomized trial were potentially unclear or ambiguous that it became necessary to add assumptions to collected questions to clarify them. Second, we discovered that independent searchers had different assumptions, so comparisons of databases were not comparing the same question.

The revised protocol that we applied was that the 3 researchers (primary care physicians with expertise in interpreting research methodology) rotated through 3 roles (A-rater, B-rater, and C-rater).

The A-rater would evaluate a question asked by a participant in the randomized trial to determine whether it met the following inclusion criteria:

1. Represents a true information need during direct patient care
2. Is clear and unambiguous
3. Is potentially answerable by medical references
4. Represents an information need likely to recur within 1 year for a typical family physician
5. Represents an information need for which fulfillment might change clinical practic

Copyright © 2005 The Annals of Family Medicine, Inc

If necessary, the A-rater could add assumptions (with documentation) to make the question meet these criteria. The A-rater would then seek the answer in DynaMed and rate the level of evidence of the answer, based on information contained within DynaMed or within 1 degree of separation (ie, by following a direct link from DynaMed and finding the answer on the page directly linked to). The level of evidence was rated from 1 (highest) to 5 (lowest) based on criteria of the Centre for Evidence-Based Medicine, (http://www.cebm.net/levels_of_evidence.asp), allowing modification of ratings for systematic reviews. Level 6 was assigned if no answer was found. Level 7 was assigned if a misleading answer was recognized and the level could be adjusted after searching other databases. The A-rater would then spend up to 30 minutes seeking the answer in the 4 key comparators—Clinical Evidence, InfoRetriever, FirstConsult (formerly PDxMD), and UpToDate—or within 1 degree of separation. Because the primary question was whether DynaMed contained an answer that met or exceeded the best answer in the other 4 sources in terms of validity, 2 shortcuts were allowed. First, if DynaMed contained a level 1 answer and any other database contained a level 1 answer, no further searching was necessary. Second, if any other database contained a better answer than DynaMed, no further searching was necessary. Otherwise, all 4 comparator databases were searched for the answer.

The A-rater would rate the level of evidence for the answers found in any comparators, note which comparator(s) had the highest level of evidence, and note whether the level of evidence in DynaMed was less than, equal to, or greater than the level of the best answer of the 4 comparators.

The B-rater would view the answers captured by the A-rater for all 5 databases and the A-rater's conclusion to the primary question (whether the level of the DynaMed answer did not meet, met, or exceeded the level of the best answer of the other databases). The B-rater would then separately rate the answers found in the 5 databases (subject to the shortcuts noted above) and had the option of doing further searches if deemed necessary. The B-rater would specifically note agreement or disagreement with the A-rater's conclusion to the primary question.

If the A-rater and B-rater agreed on all points, this result was considered the final result for the primary analysis. If there were any disagreements about the primary outcome or level of evidence ratings for individual databases, both A-rater and B-rater responses were sent to the C-rater, who made the final determination.

Because of the time delay in discovering new methods and adjusting the protocol, and the a priori belief that 400 questions would be necessary (to reach significance if DynaMed met or exceeded comparators 60% of the time), BSA became a fourth expert searcher. We set the rules so that BSA could only be an A-rater and could not serve B-rater or C-rater roles.

Because of the substantial effort involved, we consulted with our statistician, who developed a 2-stage protocol to determine whether we could confidently stop searching and reach a conclusion within the confines of the a priori confidence limits without completing 400 questions. We added a planned interim analysis at 100 questions (102 questions had actually been completed), with a decision rule to stop further analysis if DynaMed met or exceeded the other databases at least 63% of the time, representing a first-stage significance level of $\alpha = .005$.

We also analyzed whether results were dependent on who served the A-rater role by comparing percentages of answers reaching the primary outcome with each A-rater.

Because we found significant results at the interim analysis, we had time left for secondary analysis. To conduct head-to-head comparisons between individual databases, we would have to go back and complete searches not done because of shortcuts to achieve our primary analysis. There was not sufficient time left to do this for all 5 databases, so we decided to compare DynaMed with the most commonly used comparator. UpToDate was the leading comparator based on sources cited by participants enrolled in the randomized trial and our clinical experience.

DSW and ROS revisited the 102 questions and repeated searches for UpToDate if answers were not already rated as level 1 evidence. For all questions for which DSW or ROS identified new information in UpToDate, the question was then passed to the other investigator to repeat the B-rater role for this analysis. If necessary, JLH would serve the C-rater role to address disagreements, but it was not needed. We then analyzed 102 questions that each had level-of-evidence ratings for DynaMed, UpToDate, and the best of 5 databases combined to determine the distribution of levels of evidence in DynaMed, UpToDate, and the combination of 5 databases.

Copyright © 2005 The Annals of Family Medicine, Inc

Results

For the primary analysis ("Does DynaMed provide answers with validity that meets or exceeds the validity of answers provided by key comparators?"), we analyzed 102 clinical questions that were asked by participants in the randomized trial and met our inclusion criteria. Forty-five (44%) required assumptions to meet inclusion criteria.

The level of evidence for the answer found in DynaMed was equal to the best level of the comparators' answers combined for 79 (77.5%) questions, exceeded it for 10 (9.8%) questions, and was lacking for 13 (12.7%) questions.

DynaMed was therefore found to meet or exceed the comparators for 89 (87.3%) questions (95% confidence interval [CI], 80.8%-93.7%), which is significantly greater than our predefined target of 50% (*P* <.001).

Disagreements were uncommon. In 102 questions, a C-rater was needed for only 13 questions. Four of the disagreements resolved by the C-rater were between DynaMed being equal to vs DynaMed exceeding the other resources (2 determined as equal, 2 determined as exceeding), which would not change the analysis of lacking vs equal or exceeding. Five of the disagreements did not represent disagreements in question interpretation or evidence rating, but rather missing information in the initial search (the B-rater identified information not found by the A-rater) and in all cases moved DynaMed from lacking to equal as the B-rater found answers in DynaMed. Among the remaining 4 disagreements, 2 modified BSA's initial rating of DynaMed being lacking to a final rating of DynaMed being equal, 1 modified ROS's initial rating of DynaMed being equal to being lacking, and 1 did not modify DSW's initial rating of DynaMed being lacking. A sensitivity analysis using a worst-case scenario, assuming that all 8 questions for which DynaMed was initially found lacking were left that way, suggests that DynaMed met or exceeded the comparators for 82 (80.4%) questions (95% CI, 72.6%-88.2%, *P* <.001).

We also analyzed the questions for which answers were found in both DynaMed and at least 1 of the comparators, to ensure evidence quality comparisons were not skewed by questions for which level 6 (no answer found) was recorded. There were 89 questions for which answers were found in both DynaMed and at least 1 comparator, among which the level of evidence for the answer in DynaMed was equal to that of the best of the comparators combined for 73 (82.0%) questions, exceeded that of the comparators for 10 (11.2%) questions, and was lacking for 6 (6.7%) questions. DynaMed was therefore found to meet or exceed the comparators for 83 (93.3%) questions (95% CI, 88.1%-98.5%), which was significantly greater than our predefined target of 50% (*P* <.001).

Analyzing the results according to who served as the A-rater suggested possible bias when BSA was included as a searcher. The percentages of questions for which the level of evidence of DynaMed answers met or exceeded that of answers in the other 4 databases combined, by A-rater, were as follows:

Table 1. Percentage of Answers for Which DynaMed Level Met or Exceeded Level of All Comparators

A-rater	Number of Questions Rated	Initial Results	Final Results
BSA	43	90.7	95.3
JLH	5	80	80
ROS	28	75	78.6
DSW	26	76.9	84.6

Initial results are based on the A-rater's initial ratings. Final results are based on changes suggested by the B-rater that were accepted by the C-rater. These differences, however, did not differ significantly (*P* = .31 for initial results; *P* = .18 for final results).

On examining only the 59 questions for which JLH, ROS, or DSW did the initial searching, the level of evidence for the answer in DynaMed was equal to the best of that of the comparators combined for 42 (71.2%) questions, exceeded that of the comparators for 6 (10.2%) questions,

Copyright © 2005 The Annals of Family Medicine, Inc

and was lacking for 11 (18.6%) questions. DynaMed was therefore found to meet or exceed the comparators for 48 (81.4%) questions (95% CI, 71.4%-91.3%), which was still significantly greater than our predefined target of 50% (*P* <.001). BSA was able to find answers in DynaMed for 3 (27%) of the 11 questions that were rated as lacking in this analysis, and these answers were considered acceptable by JLH, ROS, and DSW (bringing the rate for which DynaMed met or exceeded the comparators up to 86.4%). This finding suggests that the differences related to BSA in the A-role were primarily due to searching proficiency rather than evidence assessment.

We conducted a secondary analysis of these 102 questions in which we repeated searches for answers in DynaMed and UpToDate, and noted the level of the best available evidence in all 5 databases.

The distribution of levels of evidence is shown in the table below.

Table 2. Number of Answers

Level of Evidence	DynaMed (n = 102)	UpToDate (n = 102)	Best of 5 Databases (n = 102)
1	50	45	55
2	12	11	14
3	1	0	1
4	3	4	3
5	23	30	23
6 (no answer)	13	12	6
7 (misleading answer)	0	0	0

The rate of achieving the best level of evidence of all 5 databases was 87.3% for DynaMed and 80.4% for UpToDate. When DynaMed and UpToDate were compared, they had the same level of evidence for 74 (72.5%) questions, the DynaMed level exceeded the UpToDate level for 17 (16.7%), and the UpToDate level exceeded the DynaMed level for 11 (10.8%) (*P* = .26). In an analysis of the 85 questions for which both DynaMed and UpToDate had answers (ie, a rating of answer validity and not scope of content), DynaMed and UpToDate had the same level of evidence for 66 (77.6%) questions, the DynaMed level exceeded the UpToDate level for 13 (15.3%), and the UpToDate level exceeded the DynaMed level for 6 (7.1%) (*P* = .11).

Interpretation and Future Research

The validity comparison shows that DynaMed meets or exceeds the validity of a substantial combination of leading clinical references. This result was found 87.3% of the time, far exceeding our feasibility criterion of 50%.

This online Supplemental Appendix is provided because, without an assessment of the validity of answers found with DynaMed, the result of increased efficiency in finding answers to questions with this database is insufficient to draw conclusions about its clinical usefulness.

This online Supplemental Appendix, extracted from the final grant report, was not prepared for formal manuscript submission, however. Given the concerns of multiple protocol changes, subjective outcomes, and the involvement of investigators with competing interests, we are seeking confirmation of these findings by independent investigators. At the time of manuscript submission, ROS and 3 independent investigators not involved in this original pilot study have developed and pilot tested a protocol for completing this study and are seeking funding support.

Copyright © 2005 The Annals of Family Medicine, Inc