Research Paper ■

Performances of 27 MEDLINE Systems Tested by Searches with Clinical Questions

R. BRIAN HAYNES, MD, PHD, CYNTHIA J. WALKER, MLS, K. ANN MCKIBBON, MLS, MARY E. JOHNSTON, BSC., ANDREW R. WILLAN, PHD

Abstract **Objective:** To compare the performances of online and compact-disc (CD-ROM) versions of the National Library of Medicine's (NLM) MEDLINE database.

Design: Analytic survey.

Setting: Health Information Research Unit, McMaster University, Hamilton, Ontario, Canada.

Intervention: Clinical questions were drawn from 18 searches originally conducted spontaneously by clinicians from wards and clinics who had used Grateful Med Version 4.0. Clinicians' search strategies were translated to meet the specific requirements of 13 online and 14 CD-ROM MEDLINE systems. A senior librarian and vendors' representatives constructed independent searches from the clinicians' questions. The librarian and clinician searches were run through each system, in command mode for the librarian and menu mode for clinicians, when available. Vendor searches were run through the vendors' own systems only.

Main Measurements: Numbers of relevant and irrelevant citations retrieved, cost (for online systems only), and time.

Results: Systems varied substantially for all searches, and for librarian and clinician searches separately, with respect to the numbers of relevant and irrelevant citations retrieved ($p < 0.001$ for both) and the cost per relevant citation ($p = 0.012$), but not with respect to the time per search. Based on combined rankings for the highest number of relevant and the lowest number of irrelevant citations retrieved, the SilverPlatter CD-ROM MEDLINE clinical journal subset performed best for librarian searches, while the PaperChase online system worked best for clinician searches. For cost per relevant citation retrieved, Dialog's Knowledge Index performed best for both librarian and clinician searches.

Conclusions: There were substantial differences in the performances of competing MEDLINE systems, and performance was affected by search strategy, which was conceived by a librarian or by clinicians.

■ **J Am Med Informatics Assoc.** 1994;1:285–295.

Affiliation of the authors: McMaster University Faculty of Health Sciences, Hamilton, Ontario, Canada.

Supported in part by the National Library of Medicine under grant R01 LM 04696 and by a National Health Scientist Award (Dr. Haynes) from the National Health Research and Development Program of Canada.

Correspondence and reprints: R. Brian Haynes, MD, PhD, McMaster University Medical Centre, Room 3H7, 1200 Main Street W., Hamilton, Ontario L8N 3Z5, Canada.

Received for publication: 12/17/93; accepted for publication: 1/20/94.

Clinical end-user searching of MEDLINE has risen dramatically during the past five years,[1] spurred by the development of "user-friendly" software, a proliferation of online and compact-disc formats, falling user charges, and advertising directed at clinicians. In 1985, before compact-disc (CD-ROM) systems became available, we conducted a study of 14 online MEDLINE routes and found considerable differences in the yields and costs of retrieval for six standardized searches run through each system.[2] By 1990, when we began the study described here, many of the

Table 1 ■

Vendors and Product Names of Online Systems That Were Compared

Vendor	Product
National Library of Medicine 8600 Rockville Pike Bethesda, MD 20892	NLM Direct Grateful Med—PC and Macintosh
BRS Information Technologies 8000 Westpark Drive McLean, VA 22102	BRS BRS After Dark BRS Colleague
DIALOG Information Services, Inc. 3460 Hillview Avenue Palo Alto, CA 94304	DIALOG DIALOG Medical Connection Knowledge Index
Personal Bibliographic Software P.O. Box 4250 Ann Arbor, MI 48106	Pro-Search—DIALOG Pro-Search—BRS
PaperChase Longwood Galleria 350 Longwood Avenue Boston, MA 02115	PaperChase
Data-Star (U.S. Office) D-S Marketing, Inc. 485 Devon Park Drive, #110 Wayne, PA 19087	Data-Star

systems had evolved and changed and some of the software from the first study had been discontinued. In addition, many new systems had emerged, so that there were six vendors offering online services through 13 different products and seven vendors offering CD-ROM products in 14 different formats.

All systems provide access to all or a subset of the MEDLINE bibliographic database prepared by the National Library of Medicine (NLM) and must meet strict NLM criteria for performance, currency, and updating. Thus, the core contents are virtually identical. Nevertheless, each system offers and advertises a number of unique features for users based on front-end software innovations and various file and pricing options.

Our interest is in the use of MEDLINE by clinicians to aid clinical decision making. We have demonstrated that clinicians with some experience in MEDLINE searching retrieved as many relevant citations as did librarians searching with the same questions, although clinicians' searches were less "precise," retrieving more citations that were irrelevant to the search question.[3] Furthermore, after basic training, clinicians acquired comparable proficiency to that of librarians by their eighth searches.[4] However, the use of MEDLINE searching by clinicians was deterred by user charges[5] and probably by other system features as well. Unfortunately, there is only one other trial of different systems of which we are aware[6]; this study found that medical-student searches on PaperChase were somewhat more productive and less expensive than those on BRS Colleague following two and a half hours of training. We are unaware of any study that has compared more than two systems.

An ideal evaluation of MEDLINE systems might involve providing access in clinical settings to all systems, randomly allocating large numbers of clinicians to use one of the systems each, and assessing the system-specific effects of MEDLINE use in terms of search performance, and effects on clinical decision making, physician performance, and patient outcome. Aside from the problems of measuring clinical effects, it is clearly not feasible to conduct a controlled clinical trial of 27 MEDLINE systems simultaneously. While a trial with two or three systems might be possible, there is little evidence to guide the selection of the systems. Thus, we elected to conduct a technical comparison of all systems, running the same searches through each system, to determine whether there were some systems that outperformed the others for retrieval of relevant citations, cost, and time per search. If a small number of systems outperformed the rest, then a controlled trial under usual clinical "field" conditions could be the next step.

In the study reported here, we attempted to simulate some aspects of clinical use of the systems by using original questions and searches generated spontaneously by clinicians who used one of the systems, *Grateful Med*, through microcomputers and modems in the wards and clinics of a teaching hospital. We also compared the clinicians' searches with searches prepared independently by a senior librarian from the clinicians' questions. In addition, system vendors were invited to send representatives to perform independent searches from the same clinical questions on their own systems.

Methods

The study methods have been described in detail elsewhere.[7] All MEDLINE products available at the beginning of the study in 1990 were included (Tables 1 and 2). Each product from a given vendor was tested as a separate system. For example, Grateful Med for the PC and Macintosh counted as two products, as did each subscription type of CD-ROM (full MEDLINE database, subset, monthly, or quarterly, as offered by the vendor).

Each system was tested with 18 search questions drawn from those posed by clinicians in the control

group of a randomized controlled trial of an intervention to improve the use of MEDLINE in hospital wards and clinics.[4] Clinicians, who had received basic instruction in the nature of MEDLINE and use of Grateful Med Version 4.0, could search on any of 16 computers throughout the hospital, each equipped with Grateful Med. Searches performed at the midpoint of the study were eligible for selection when they were about a patient problem and had a complete and understandable question. From the eligible searches, we selected three by random process from each of six clinical departments, with no more than two questions from the same searcher. The questions were stratified to obtain six questions on therapy, six on diagnosis, and six on prognosis for a total of 18 questions. Two additional questions were selected for practice runs on each system. Both the clinicians' original search strategies for each question and the librarian's strategy based on the question were run through each system.

All vendors were sent a copy of the protocol and were invited to send representatives to the study site at McMaster University to run the 12 therapy and diagnosis questions through their own systems. Reimbursement for travel expenses was offered.

The senior study librarian and system vendors constructed their searches using the following information: complete search questions, patient age and gender, and clinical department in which the search originated. The librarian and vendors were blinded to the clinician's original search strategy and to each other's searches. Librarian searches were intentionally formulated to retrieve a few highly relevant citations to answer the clinical search questions, with retrieval of a minimal number of irrelevant citations, and were not intended to maximize recall of relevant citations. Thus, librarian strategies were limited by the medical subject heading (MeSH) "human" and to English-language articles, and most included one or more methodologic MeSH terms such as "randomized controlled trial (pt)" for questions about treatments or "sensitivity and specificity" for studies about diagnosis. Vendor's representatives were not given any instructions about what strategies or limits to use for searches. The librarian and vendors had access to MeSHs for 1991[8] and a medical dictionary. The librarian could sign into the Elhill service at the NLM directly (ELHILL MEDLINE), and vendors could sign into their own systems, up to two times to verify that the search strategy was likely to be productive before committing to it.

The librarian and vendors were not permitted to consult the original searchers or other health professionals about the search questions. This approach was used to avoid three problems: original searchers refining their questions after doing their own searches; original searchers incorporating what they had encountered during their searches; and other clinicians providing advice to the librarian or vendors that would improve their success at finding relevant material. This approach also put the librarian and vendors at a slight disadvantage compared with the traditional methods of mediated searches, in which the intermediary clarifies the end-user's question before running the search. However, the effect of this on the comparison of librarian and clinician searches was felt to be minimal as only questions that we judged to be "complete and understandable" were eligible for selection for the study. In any event, this approach would affect only clinician–librarian comparisons, not system–system comparisons or librarian–vendor comparisons.

Clinicians' spontaneous, self-conducted searches using Grateful Med Version 4.0 during a randomized controlled trial[4] were extracted by a research assistant, who worked independently of the study librarian and vendors, from computerized keystroke records. If the search comprised more than one attempt, the last strategy producing citations was selected. Cross-references and text words were handled as follows: Grateful Med Version 4.0 accepted cross-ref-

Table 2 ■

Vendors and Product Names of CD-ROM Systems That Were Compared

Vendor	Product
Cambridge Scientific Abstracts 7200 Wisconsin Avenue Bethesda, MD 20814	Compact Cambridge—monthly & quarterly
CD Plus 333 Seventh Avenue, 6th Floor New York, NY 10001	CD Plus
SilverPlatter Information, Inc. 1 Newton Executive Park Newton Lower Falls, MA 02162	SilverPlatter—unabridged & subset
EBSCO Electronic Information 461 Boston Road, Unit 3D Topsfield, MA 01983	EBSCO CD-ROM—unabridged & subset
Healthcare Information Services 2335 American River Road, #307 Sacramento, CA 95825	BiblioMed BiblioMed Professional Test Version
DIALOG Information Services, Inc. 3460 Hillview Avenue Palo Alto, CA 94304	DIALOG OnDisc—unabridged & subset
ARIES Systems Corporation 1 Dundee Park Andover, MA 01810	ARIES Knowledge Finder—monthly, quarterly, & subset

erences but searched the appropriate MeSH term, even though the latter did not appear on the software's form screen. For example, if "cancer (xr)" was chosen by the clinical searcher from the MeSHs, "cancer (mh)" would appear on the form screen but "neoplasms (mh)" would be searched online. Therefore, when translating the strategies into the other systems' search protocols, the term that was searched was used, not the term that the searcher selected when it was a cross-reference. Words and terms that were typed into Grateful Med and not chosen from the MeSHs within Grateful Med were taken to mean any occurrence of that word or phrase. For example, if "cholera" was typed on a subject line in Grateful Med and not chosen from the MeSHs, it would be searched as "cholera (mh) or cholera (tw)." The translation into other systems, therefore, would be cholera as both subject heading and text word.

Translations were made as close to the original strategy as each system would allow. Exceptions were made for systems that could not accommodate certain ELHILL MEDLINE search features. If explosions were not allowed by a system, specific terms under an explodable term were OR'd if there were fewer than seven terms. If there were more than seven terms, the first seven were OR'd and the rest omitted. Systems not permitting bald subheadings were searched with attached subheadings. If the attached subheading was not applicable to that term, the subheading was searched as an AND'd text word and, if possible, as a subject heading.

Products from the same vendor were grouped and randomized to the two study librarians. Translations were done by the librarians in a prescribed order, according to the individual system's requirements. The librarian searches were translated into a given system, followed by the clinician searches. If a system had both a menu and a command mode, the librarian strategies were translated into the command mode and the clinician strategies were translated into the menu mode on the grounds that these were the most likely modes of use for each type of searcher.

For all searches, the MEDLINE file time period for searching was standardized. Two clinician search strategies generated so much retrieval for the standardized time period that they had to be interrupted. These both included text words OR'd together that logically should have been AND'd. The retrieval for each was restricted by date, if the system allowed, from October 1990 to spring 1991. If it was not possible to restrict to that time period, the first 100 citations were downloaded and numbers were extrapolated to the full time period for analysis.

Some systems required adjustments to run the strategies as they were written. In Grateful Med, some librarian strategies could not be performed within the confines of the form screen and the searcher had to "take control," using the system's command language for some of the strategy. For example, in one search, three subheadings had to be combined in a complex Boolean OR statement to avoid a "stores postings overflow" message from MEDLINE. Command searching was also necessary when combining the Abridged Index Medicus (AIM) subset with a subheading. EBSCO required two searches to perform the strategy with the AIM subset. Furthermore, two clinician searches could not be done in the menu mode on EBSCO because its menu mode only allows three AND statements.

ARIES Knowledge Finder uses probabilistic matching of search terms instead of the Boolean search approach used by other systems. Although ARIES has all the regular MEDLINE features, such as MeSHs, subheadings, and explodes, it was sometimes not feasible to search exactly as a strategy was originally written. ARIES also has search controls (namely, a relevance filter, a limit control on the number of documents retrieved, and a word variants control) that can be adjusted according to the searcher's wishes. In order to emulate the original search strategy as closely as possible, the quantity was emphasized with the relevance filter, the maximum number of documents was retrieved, and the word variants control was turned off. Some of the librarian strategies involved ORing publication types and check tags with regular MeSH terms. In ARIES, these term types could not be OR'd with MeSH terms, so the strategy was performed in two parts to accommodate those operations.

All search strategies, both clinician and librarian, were run through each system by one of the two study librarians, the librarians being randomly assigned to the systems, and the searches were timed during running and downloading. In addition, vendors ran their own searches, based on the clinicians' questions, through their own systems. Upon completion of the searches, the unique identifiers of all captured citations were entered into a single system and the citation, abstract, and MeSH terms were printed out so that all the citations would look identical. For each search question, citations retrieved by the searchers from all systems were pooled, placed in random order, and given to a clinical reviewer. This permitted clinical reviewers to assess retrieved citations for relevance without knowledge of which system produced them. Reviewers rated each citation on a seven-point scale, where 7 was directly relevant and 1 was

definitely not relevant.[3] For most analyses, citations scoring 5 or higher were treated as "relevant" and those scoring 1 to 4 were treated as "irrelevant." This rating method has clearly distinguished between inexperienced and experienced searchers in previous studies.[3,4]

Data were collected for each search session on prepared worksheets. We recorded search construction for librarian, vendor, and clinician searches, search translation into each system, and the times and costs for all search sessions. These data were keyed into a database (PARADOX Version 3.5, Borland, Scotts Valley, CA). The output from each search session was captured and data including the eight-digit unique identifier of each citation, the system number, the search question number, and the source of the strategy (librarian, clinician, or vendor) were extracted using a purpose-written UNIX AWK script and uploaded into PARADOX tables. Each system was assigned a code so that analyses could be conducted that were blinded to the name of the vendor.

Main outcomes were pre-defined as the average number of relevant and irrelevant citations retrieved for all searches run through a given system. Recall was calculated as the number of relevant citations retrieved for a search divided by the number of relevant citations retrieved for that search question across all systems. Also captured were cost per search and cost per relevant citation for online systems, and time per search, processing time per search, and time per relevant citation for all systems. These values were keyed into the PARADOX tables for each search session, separately for clinician, librarian, and vendor searches, and grouped by online or CD-ROM systems with results for each system averaged across the 18 search questions.

The performances of systems for average values of outcome measurements were compared by analysis of variance (ANOVA). To stabilize variances, the average number of relevant citations was converted to its square root for each system and the average number of irrelevant citations was converted to its fourth

Table 3 ■

Ranks, Based on Retrieval of Relevant and Irrelevant Citations, for All Systems for Clinician Searches

Systems Ordered by Combined Rank (Sum of Ranks for Relevant and Irrelevant Citations)	Relevant Citations Retrieved			Irrelevant Citations Retrieved		
	Mean No.	SE	Rank	Mean No.	SE	Rank
1. PaperChase	7.4	4.03	2	38.5	25.67	10
2. Dialog OnDisc unabridged (CD)	8.4	5.38	1	51.7	42.44	13
3. Compact Cambridge monthly (CD)	6.7	3.64	8.5	51.0	38.83	12
4. Grateful Med Macintosh	7.0	4.18	5	54.8	40.10	16.5
5. ARIES unabridged monthly (CD)	5.4	3.52	18	19.6	8.33	4
6. Knowledge Index	6.8	3.97	6.5	54.9	40.29	18
7. ARIES unabridged quarterly (CD)	3.3	1.55	23	13.4	5.62	2
8. Bibliomed Professional Test (CD)	5.1	2.70	19	31.3	19.71	7
10.5 ARIES subset (CD)	1.1	0.57	27	4.9	2.31	1
10.5 Compact Cambridge quarterly (CD)	5.9	3.46	17	49.2	36.66	11
10.5 Dialog OnDisc subset (CD)	2.3	1.39	25	16.7	13.76	3
10.5 Grateful Med PC	7.1	3.86	4	55.9	41.78	24
14. BRS After Dark	6.6	3.78	12	54.8	41.74	16.5
14. Dialog	6.8	3.92	6.5	55.6	40.99	22
14. Dialog Medical Connection	6.7	3.92	8.5	55.2	41.04	20
17. Bibliomed (CD)	2.6	1.13	24	21.9	17.58	5
17. CD Plus (CD)	4.9	3.20	21	32.2	11.73	8
17. EBSCO unabridged (CD)	7.3	3.47	3	61.0	41.02	26
19.5 Pro-Search—Dialog	6.4	3.49	15	54.1	40.19	15
19.5 SilverPlatter unabridged (CD)	6.3	3.94	16	53.4	42.15	14
21.5 DataStar	6.6	3.76	12	55.0	41.12	19
21.5 EBSCO subset	3.9	1.75	22	36.1	25.82	9
23. SilverPlatter subset (CD)	2.2	1.09	26	24.7	19.71	6
24. BRS	6.6	3.78	12	55.3	41.61	21
25. ELHILL	6.6	3.84	12	55.7	41.62	23
26. Pro-Search—BRS	6.6	3.78	12	56.2	41.58	25
27. BRS Colleague	5.0	3.31	20	64.9	51.98	27

root. Additional ANOVAs were done for search time and cost per relevant citation comparing each system with ELHILL MEDLINE (that is, MEDLINE searched in command mode via ELHILL at the NLM). If there were statistically significant differences in performance for individual outcomes among all systems (e.g., for average number of relevant citations retrieved), pairwise comparisons were made with ELHILL MEDLINE. ELHILL MEDLINE was chosen as the reference system for pairwise comparisons because it fared best in the previous evaluation study and because its database is the source for all the other systems.[2] McNemar chi-square tests with Yates' correction were used for matched comparisons of librarian and clinician searches for higher yield for relevant and irrelevant citations. All p values are two-tailed.

Results

System Details

The 27 systems that were compared are described in Tables 1 and 2. There were 13 online and 14 CD-ROM products, including 4 subsets on CD-ROM, offered by 12 vendors, with 9 vendors providing more than one product. Only one company, Dialog, offered both online and CD-ROM access. Subsequent to the study, in 1993, Compact Cambridge was taken over and discontinued by SilverPlatter.

Overall System Performance

There were highly statistically significant differences among the 27 systems for all searches (clinician and librarian combined) for a number of outcome measures, including the average number of relevant citations retrieved per search (range, 0.8 to 7.4; ANOVA F value = 4.33 on 26 degrees of freedom, $p < 0.0001$), the average number of irrelevant citations retrieved (range, 1.1 to 64.9; F value = 5.78, $p < 0.0001$), and the average cost per relevant citation (determined for online systems only, for searches that yielded at least one relevant citation) (range, \$0.62 to \$3.71; F value = 3.99, $p = 0.012$). The same levels of statistical significance were observed for these measures when performance was assessed for clini-

Table 4 ■

Ranks, Based on Retrieval of Relevant and Irrelevant Citations, for All Systems for Librarian Searches

Systems Ordered by Combined Rank (Sum of Ranks for Relevant and Irrelevant Citations)	Relevant Citations Retrieved			Irrelevant Citations Retrieved		
	Mean No.	SE	Rank	Mean No.	SE	Rank
1. SilverPlatter subset (CD)	3.5	2.70	5	2.9	1.29	17
2.5 Dialog	2.1	0.66	14	2.6	0.87	9.5
2.5 Pro-Search—Dialog	2.1	0.66	14	2.6	0.87	9.5
4. SilverPlatter unabridged (CD)	6.4	4.53	1	5.5	1.97	23
5. BiblioMed (CD)	1.2	0.53	24	1.1	0.46	1
6. Knowledge Index	3.6	1.79	4	5.0	2.40	22
8. ARIES unabridged quarterly (CD)	3.2	1.26	6	4.2	2.04	20.5
8. Data-Star	2.3	0.84	9.5	2.9	0.87	17
8. Grateful Med—Macintosh	2.3	0.66	9.5	2.9	0.87	17
10. BiblioMed Professional test (CD)	1.7	0.87	21	2.2	0.84	6
13. BRS	2.1	0.66	14	2.7	0.88	13.5
13. BRS After Dark	1.9	0.66	18	2.6	0.89	9.5
13. BRS Colleague	2.1	0.70	14	2.7	0.89	13.5
13. Dialog Medical Connection	2.1	0.66	14	2.7	0.88	13.5
13. ELHILL	2.1	0.65	14	2.7	0.88	13.5
16.5 ARIES unabridged monthly (CD)	5.1	2.51	3	7.8	3.38	25
16.5 ARIES subset (CD)	1.3	0.55	23	2.1	0.94	5
18. Dialog OnDisc subset (CD)	0.8	0.35	26	1.3	0.70	2
20.5 CD Plus (CD)	1.1	0.37	25	1.8	0.85	4
20.5 EBSCO subset (CD)	1.4	0.56	22	2.3	0.81	7
20.5 Grateful Med PC	1.8	0.63	19.5	2.6	0.90	9.5
20.5 PaperChase	5.5	2.73	2	9.0	5.45	27
23. Dialog OnDisc unabridged (CD)	0.8	0.35	26	1.4	0.69	3
24. Compact Cambridge quarterly (CD)	2.9	1.18	8	7.6	4.90	24
25. Compact Cambridge monthly (CD)	3.0	1.19	7	8.1	5.23	26
26. Pro-Search—BRS	2.1	0.60	14	4.2	1.61	20.5
27. EBSCO unabridged (CD)	1.8	0.64	19.5	3.6	1.07	19

cian and librarian searches separately. When all searches were pooled, including those that did not yield a relevant citation, the cost per relevant citation ranged from $0.81 to $4.07 for clinician searches and from $0.46 to $2.09 for librarian searches. Time per relevant citation retrieved did not differ across systems (F value = 1.19, $p = 0.333$), although the range was quite wide at 1.60 to 4.39 minutes per relevant citation for the online systems and 1.09 to 6.40 minutes for the CD-ROM systems. There were no statistically significant differences in the proportion of searches producing no relevant citations, with the range extending from 0.33 to 0.44 for online systems and from 0.33 to 0.67 for CD-ROM systems.

Individual System Comparisons

In Tables 3, 4, and 5, systems are rank-ordered, with lower ranks for better performance, from three different perspectives for outcome measures for which there were statistically significant overall performance differences. As shown in Tables 3 and 4, the top-ranked systems differed for clinician and librarian searches and for CD-ROM and online systems. Judged by the number of relevant citations retrieved, Dialog OnDisc unabridged MEDLINE was best for clinician searches, while SilverPlatter unabridged MEDLINE version on CD-ROM performed best for librarian searches. Combining rankings for the highest number of relevant and the lowest number of irrelevant citations retrieved, the PaperChase online service performed best for clinician searches, while SilverPlatter CD-ROM subset performed best for librarian searches.

Cost per relevant citation was measured for online systems and differed significantly across these systems. As shown in Table 5, Dialog's Knowledge Index charged the least per relevant citation for both clinician and librarian searches. When the ranks for relevant and irrelevant citations retrieved (Tables 3 and 4) are added to those for cost per relevant citation (Table 5), PaperChase was best for clinician searches and the Dialog online MEDLINE service accessed through Pro-Search performed best for librarian searches.

For the CD-ROM systems, cost per relevant citation retrieved could not be measured directly because there are no online charges. Based on combined ranks for relevant and irrelevant citations retrieved from CD-ROM systems alone, ARIES Knowledge Finder full MEDLINE monthly version worked best for clinician searches, while the SilverPlatter clinical subset worked best for librarian searches (more details available on request). These findings apply within only the CD-ROM systems and appear to be somewhat different from the results shown in Table 3 because of the intervening ranks of the online systems.

Comparisons with ELHILL MEDLINE

In pairwise comparisons with ELHILL MEDLINE for clinician searches, adjusting for the number of comparisons, the SilverPlatter clinical subset, the Dialog OnDisc clinical subset, and the ARIES Knowledge Finder clinical subset had significantly fewer relevant and irrelevant citations than did ELHILL MEDLINE, and BiblioMed had significantly fewer irrelevant citations ($p < 0.05$). For librarian searches, the Dialog OnDisc clinical subset had fewer irrelevant citations ($p < 0.05$). All of these systems are CD-ROM systems. To compare search costs for online systems

Table 5 ■

Ranks and Costs per Relevant Citation for Online Systems for Clinician and Librarian Searches

Systems Ordered by Combined Rank (Sum of Ranks for Clinician and Librarian Searches)	Clinician Searches		Librarian Searches	
	Cost per Relevant Citation	Rank	Cost per Relevant Citation	Rank
1. Knowledge Index	1.02	1	0.62	1
2. BRS After Dark	1.33	2	0.89	2
3. Data-Star	1.43	3	1.14	4
4. Dialog	1.99	4	1.44	6
5.5 Pro-Search—Dialog	2.38	8	1.00	3
5.5 Dialog Medical Connection	2.09	6	1.35	5
7. BRS Colleague	2.04	5	1.70	9
8.5 Grateful Med—Macintosh	2.48	10	1.45	7
8.5 Grateful Med—PC	2.45	9	1.53	8
10. PaperChase	2.11	7	2.22	11.5
11. ELHILL	2.92	11	1.88	10
12. Pro-Search—BRS	3.71	12	2.22	11.5

with costs for ELHILL MEDLINE, clinician and librarian cost data were combined. Knowledge Index charged less per relevant citation than did ELHILL MEDLINE ($p < 0.05$).

Clinician and Librarian Searches

Although not the main focus of the study, there were differences in clinician and librarian searches in addition to those in system performance rankings (see above). Librarian searches were intentionally designed to be highly focused, retrieving a few relevant references and fewer irrelevant references, and this approach was evident in the results. Clinician search strategies retrieved more relevant citations for all online systems (McNemar chi-square, 1 degree of freedom, $p = 0.009$) but also retrieved more irrelevant citations for all online systems ($p = 0.009$), with clinician searches costing more than librarian searches per relevant citation on online systems (difference = \$1.12, $p < 0.0001$). However, the total searching time per relevant citation was not different (2.3 vs 2.6 minutes, respectively, $p = 0.12$), presumably because the clinician searches were less complex. For CD-ROM systems, the findings were similar but less marked. The clinician searches had somewhat higher retrieval of relevant citations for all but three of the 14 systems ($p = 0.061$) and had a higher number of irrelevant references for all CD-ROM systems ($p = 0.0005$). In contrast to online systems, clinician searches took less time to run than did librarian searches on CD-ROM systems, the average difference being 1.88 minutes ($p = 0.0008$). The differences between clinician and librarian searches were particularly marked for irrelevant retrieval, for both online and CD-ROM systems, with clinician searches having from 5 to 65 irrelevant retrievals per search and most librarian strategies retrieving from one to no irrelevant article per search. Thus, clinician searches were generally much less precise than librarian searches.

Librarian and Vendor Searches

Thirteen vendors were invited to McMaster University to perform searches on their own systems. All but one vendor accepted (PaperChase declined). Because of time constraints, most vendors with more than one system ran searches through only one of their products. Thus, librarian and vendor searches were compared for 12 searches for each of six online systems and seven CD-ROM systems. Overall, vendors' searches retrieved higher numbers of relevant citations (ANOVA, $p = 0.001$) and higher numbers of irrelevant citations ($p = 0.002$). Thus, vendors' search strategies resembled clinicians' search strategies in terms of yield.

Discussion

We found highly statistically significant differences in MEDLINE systems for average number of relevant citations retrieved and average number of irrelevant citations retrieved for all systems combined, for online systems combined, and for librarian searches compared with both clinician searches and vendor searches. Clinician search strategies were not compared directly with vendor strategies but appeared to be similar in retrieving higher numbers of both relevant and irrelevant citations than did librarian searches.

In ranking systems by performance, no one system performed consistently best for both relevant and irrelevant retrieval (and cost for the online systems) and there was no type of system (online or CD-ROM) that performed consistently best. For clinician searches, an online system, PaperChase, performed best, while for librarian searches, a CD-ROM system, the SilverPlatter clinical subset, performed best overall. However, the differences between systems ranked closely together were often very small and individual system features may be more important to users than the overall rankings reported in the tables.

In pairwise comparisons with the ELHILL MEDLINE reference system, four CD-ROM systems differed significantly for clinician searches, with three retrieving both fewer relevant and irrelevant citations than did ELHILL MEDLINE and one retrieving fewer irrelevant citations. Only one system differed significantly in terms of relevance for librarian searches. Only one system differed significantly for cost, and there were no significant differences for time. In the rankings, ELHILL MEDLINE was ranked thirteenth for relevant and irrelevant citation retrieval combined for librarian searches and twenty-fifth for clinician searches on the combined measure. Grateful Med was developed with the intention of assisting nonlibrarian end-users to circumvent the specialized command language of MEDLARS. In the rankings, Grateful Med for PCs performed better than ELHILL MEDLINE for clinician searches and less well for librarian searches, while Grateful Med for Macintosh computers outperformed both ELHILL MEDLINE and the PC version of Grateful Med.

There are some limitations of our research that must be borne in mind when interpreting the findings. We did not attempt to compare clinicians' and librarians' searches directly. Rather, the search questions and clinicians' search strategies were drawn from spontaneous searches by clinicians that were all originally conducted on Grateful Med Version 4.0. This may

have imposed searching patterns that affected search performance on other systems. However, these search strategies worked well on CD-ROM systems, with five of these systems placing in the top 10 combined rankings. This may have been because of the simplicity of the general style of clinician searches, which were characterized by a limited number of concepts AND'ed together. This is consistent with Cahan's finding that 40% of end-user searches were based on single search statements.[9] By contrast, the librarian searches often included explosions and pre-explosions. Other studies have shown that librarians frequently make use of such features.[10] In our investigation, for combined rankings on relevant and irrelevant retrieval frequency, the top 10 rankings for librarian searches were shared equally by CD-ROM and online systems, indicating that neither approach can claim superiority.

Only one librarian was involved in developing the basic search strategies for the librarian searches and these strategies reflected her chosen style, namely, tightly focused searches for a small number of clinically relevant articles. In previous investigations, this librarian's performance has been shown to be indistinguishable in yield of relevant and irrelevant citations from other senior librarians with extensive search experience.[7] However, the systems performed differently with other styles, as documented by the results for both the vendors' representatives, many of whom were librarians themselves, and the clinicians.

Familiarity with a given system may be a key factor in how search strategies are formulated and thus in how well a system performs. The study librarians were most familiar with the command mode of ELHILL MEDLINE and the Grateful Med PC software and this could have affected the findings. However, these systems did not seem to be favored by the findings, perhaps because of the extensive efforts in the study to neutralize any such effects. The librarians had had extensive search and teaching experience with several other systems, had studied the documentation for each system's requirements, and had run practice searches through each system before the formal system evaluation searches began. Furthermore, significant and substantial differences in performance emerged, both when all searches were combined and when clinician and librarian searches were considered separately. Thus, differences in performance cannot be explained solely on the basis of familiarity with the systems and it is unlikely that this was a major factor.

With 27 systems being compared, it is difficult to identify systematic features that explain the differences in performance. All systems have the same core citation database or subsets of it and most systems use a search program that is similar to STAIRS used by ELHILL, so it is not surprising that performances are similar in several ways, particularly for the librarian searches that were developed by one person. The clinician searches, however, were from several end-users and lend themselves to some speculation about differences in system performance. PaperChase uses some artificial-intelligence features to simplify the user interface for clinicians and enhance search yield and this may be the reason for its success for the clinician searches. Among the CD-ROM systems, ARIES Knowledge Finder stands out for its fuzzy-matching and relevance-ranking features and also performed well for clinical searches.

We used the current version of each system when the study was conducted in 1990 and 1991, but the systems continue to evolve. One of the CD-ROM systems we tested, Compact Cambridge, has been bought and discontinued by another, SilverPlatter. The NLM also continues to perfect article indexing, a notable improvement being the introduction of publication types that have increased the ability of searches to retrieve clinically important citations on, for example, randomized controlled trials.[11] These changes can be expected to affect the performance of all systems to the extent that they provide users access to the new indexing features. In addition, many of the systems have reduced their user charges and the cost comparisons may no longer be representative.

Searching styles of end-users and librarians have been compared previously but the findings are not consistent. Sullivan et al. found that end-users captured as many relevant citations as did librarians with smaller overall retrievals.[12] Our own previous research revealed higher rates of recall and precision for librarians over clinicians who were inexperienced searchers, but these clinicians did retrieve some unique relevant citations not captured by either librarians or experienced clinician searchers.[7] Furthermore, experienced clinical searchers retrieved as many relevant citations as did librarians, although librarians retained the lead in smaller numbers of irrelevant citations retrieved.[2,7]

The large number of online and CD-ROM systems and the changes in indexing and formats make it difficult to perform comprehensive assessments with durable conclusions, and that was not our purpose. We wished to determine whether the systems performed similarly when presented with a set of searches drawn from clinical practice and also when a set of tightly focused librarian searches were done. This might be termed "field testing" and does not con-

stitute a rigorous clinical trial with physician performance and patient outcomes as endpoints. The results are clear: the systems do not perform the same way. Additional studies are therefore warranted to determine which systems perform best in direct end-user tests in specific settings, preferably using randomized trial designs, limiting the comparisons for practical purposes to the best performing systems from our investigation. This staged approach to evaluation is in keeping with recent recommendations for testing informatics innovations.[13] In lieu of such advanced studies, vendors may be able to use the information from our study to enhance their systems and users may find the information helpful in making decisions about which systems to access.

References ■

1. Dr. Lindberg reports to the Congress: 1991—The year of outreach. Gratefully Yours. March/April 1992:4–5.
2. Haynes RB, McKibbon KA, Fitzgerald D, et al. Computer searching of the medical literature: an evaluation of MEDLINE searching systems. Ann Intern Med. 1985;103:812–6.
3. Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online access to MEDLINE in clinical settings: a study of use and usefulness. Ann Intern Med. 1990;112: 78–84.
4. Haynes RB, Johnston ME, McKibbon KA, Walker CJ, Willan AR. A randomized controlled trial of a program to enhance clinical use of MEDLINE. Online J Curr Clin Trials [serial online]. 1993;May 11;Doc No. 56.
5. Haynes RB, Ramsden MF, McKibbon KA, Walker CJ, Johnston M. Online access to MEDLINE in clinical settings: impact of introducing user fees. Bull Med Libr Assoc. 1991;79:377–81.
6. Porter D, Wigton RS, Reidelbach MA, Bleich HL, Slack WV. Self-service computerized bibliographic retrieval: a comparison of Colleague and PaperChase, programs that search the MEDLINE data base. Comput Biomed Res. 1988;21:488–501.
7. Walker CJ, McKibbon KA, Haynes RB, Johnston ME. Performance appraisal of online MEDLINE access routes. Proc Annu Symp Comput Appl Med Care. 1992;16:483–7.
8. National Library of Medicine. Medical Subject Headings, Annotated Alphabetic List, 1991. Distributed by the National Technical Information Service, U.S. Department of Commerce. PB-91-1000008.
9. Cahan MA. GRATEFUL MED: a tool for studying searching behavior. Med Ref Serv Q. 1989;8:61–79.
10. McKibbon KA, Haynes RB, Walker Dilks CJ, et al. How good are clinical MEDLINE searches? A comparative study of clinical end-user searches and librarian searches. Comput Biomed Res. 1990;23:583–93.
11. Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. Proc Annu Symp Comput Appl Med Care. 1993;17:601–5.
12. Sullivan MV, Borgman CL, Wippert D. End-users, mediated searches, and front-end assistance programs on Dialog: a comparison of learning, performance, and satisfaction. J Am Soc Info Sci. 1990;41:27–42.
13. 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;1:28–34.

APPENDIX A

Vendors' Comments

Vendors were provided with the coded results for comment. Substantive comments about the study that were not addressed in the text of the report are briefly recorded here. Our replies before the code was broken appear in square brackets and those after the code was broken appear in curly brackets. We have not included claims for improvements in searching capabilities or pricing since the time of the study as all vendors have made changes but evidence of the effects of these changes on comparisons between systems is not available.

ARIES

1. Knowledge Finder searches were performed with quantity emphasized, maximum documents, and word variants off. The first two modifications would have the effect of producing many more irrelevant documents, while (typically) not generating additional relevant documents. The third modification reduces the likelihood of retrieving certain relevant documents. Knowledge Finder's performance would be expected to be significantly affected by these settings. Knowledge Finder is typically used with standard settings: balanced quality/relevance, 100-document maximum, and word variants on. {Despite these considerations, Knowledge Finder performed well in comparison with other systems.}

2. The ranking statistics in the report do not include time spent to achieve search results. [There were no significant differences in time across the systems so we did not report these in detail.]

CD PLUS

1. The masked nature of the study and the lack of specific information about the search strategies makes it impossible to comment on whether the data were presented fairly. [Vendors were presented with only coded data to ensure that their comments about data presentation were not biased by the performance of their own systems. The search strategies are available on request but are not particularly relevant to the objective of the study as the same searches were put through each system.]

2. We are concerned that the study was prejudiced because Grateful Med was used to determine the sample clinical searches. [Clinician searches were mostly quite simple and were modified for the special requirements for each system. The librarian searches were developed independently, using command language. ELHILL MEDLINE did not appear to be favored in the comparisons with other systems.]

Data-Star

1. You focus in the Discussion section on the comparison of clinician and librarian searches, less on the vendors' searches. [True. Vendors' searches were run only through their own systems, and thus could not be used to compare

one system with another. Compared with librarian searches, vendors' searches retrieved more relevant references and also more irrelevant references, thus resembling the less tightly focused searches of clinicians. Another study would be required to do a more detailed evaluation of vendors' searches.]

PaperChase

1. Clinicians' searches were taken from one system. To the extent that certain questions can be answered more easily with one system, the study was biased. [PaperChase states that it makes searching easier for clinicians. To the extent that this is true, this should favor PaperChase.] {The results show that PaperChase actually performed better than Grateful Med and the other systems for clinicians' search strategies despite whatever handicap might have been imposed by the origin of the searches.}

2. Although the manuscript states that the databases for all systems were "virtually identical," PaperChase included HEALTH in the same file as MEDLINE. [The HEALTH file (articles on health planning and administration) is relatively small and probably would not have contributed to the searches in the study, all of which dealt with strictly clinical problems.]

3. The manuscript discusses search techniques that are appropriate for programs that resemble STAIRS [the ELHILL program] but inappropriate for programs that do not. For example, the filters human and English were used. In the case of PaperChase, artificial intelligence handles these and other such matters automatically. If these terms had been used in PaperChase searches, the result would have been increased typing, slower search speed, and higher cost. {PaperChase seems to have done well for clinicians' searches just the same and retained its overall number 1 ranking for clinician searches even when cost per relevant citation was incorporated into the ranking.}

4. Although the searches were from clinicians, these searches were not performed on the systems by clinicians and therefore do not have any bearing on the question of which system is best for clinicians. [Although the link to clinicians is one step removed, the original searches were formulated by clinicians and their searches ran better through some systems than others. Having clinicians run their searches through more than one system themselves is logistically complicated and methodologically questionable. Randomly allocating clinicians to 27 different systems is unfeasible.]

5. The study was funded by one of the vendors, yet the manuscript describes no effort to avoid bias in favor of that vendor's system. [The project was funded by the NLM after excellent ratings from independent peer review. The NLM is the only peer-review funding agency among the vendor group. Retrievals from all systems were coded and judged without knowledge of the system of origin, and the results were analyzed without breaking the code. Vendors were given the opportunity to visit and run searches through their own systems; all but one, PaperChase, did so. All vendors reviewed the manuscript before the code was broken.] {Given the results of the study, PaperChase's reservations seem unfounded.}