INTRODUCTION

In January 1966 the National Library of Medicine (NLM) embarked upon the detailed planning of a test program to evaluate the performance of MEDLARS (Medical Literature Analysis and Retrieval System). In December 1965, the writer had been recruited by the Library to fill the new position of Information Systems Evaluator, thus enabling the evaluation to be conducted in a completely impartial manner by someone who had in no way been concerned with either design or operation of the MEDLARS system. This spirit of impartial analysis has been maintained by the Evaluator throughout the evaluation program.

In addition, the Director of the National Library of Medicine appointed a MEDLARS Evaluation Advisory Committee, to review the design and execution of the test program, and the analysis and presentation of the test results. This committee, for whose advice and criticism the writer is deeply indebted, has consisted of the following members:

Charles J. Austin, Director of Computer Services and Assistant Professor
University of Colorado Medical Center, Denver, Colorado
Dr. Julian Bigelow, Permanent Member, The Institute for Advanced Study
Princeton, New Jersey
Cyril W. Cleverdon, Librarian, College of Aeronautics, Cranfield, England
W. D. Climenson, Deputy Director of Computer Services, Central Intelligence Agency
Dr. Eugene K. Harris, Chief, Laboratory of Applied Studies, Division of Computer Research and Technology, National Institutes of Health
Dr. Calvin Mooers, President, Rockford Research Institute Inc.
Cambridge, Massachusetts

The methodology and findings of this study were fully endorsed by this committee at its final meeting on January 15-16, 1968.

Cyril Cleverdon has acted as special consultant to the Library on this project. His assistance has been invaluable, particularly in the design and analysis phases of the program.

The author is also deeply grateful for the willing help given to him by the library and information staff of the 20 organizations participating in this evaluation program.

PART 1

DESIGN AND EXECUTION OF THE EVALUATION PROGRAM

MEDLARS: GENERAL BACKGROUND

The Medical Literature Analysis and Retrieval System has been discussed in detail elsewhere.[1] Only the most salient characteristics will be described here.

MEDLARS is a multipurpose system, a prime purpose being the production of Index Medicus and other recurring bibliographies. However, the present study has concentrated on the evaluation of the demand search function (i.e., the conduct of retrospective literature searches in response to specific demands). The base of the retrospective search module consists of more than half a million citations to journal articles, in the biomedical field, input to the January 1964 and subsequent issues of the monthly Index Medicus. This data base is presently growing at the approximate rate of 200,000 citations annually. Journal articles, of which roughly 45% are in languages other than English, are indexed at an average level of 6.7 terms per item, using a controlled vocabulary of Medical Subject Headings (MeSH). Over three thousand demand searches are processed annually at the National Library of Medicine, additional searches being handled at regional MEDLARS centers in the United States, in the United Kingdom and in Sweden.

Approximately 2400 scientific journals are indexed regularly. About one third of these are indexed exhaustively ("depth journals") at an average of 10 terms per article, and the remainder are indexed less exhaustively ("non-depth journals") at an average of slightly under four terms per article.

MeSH consists of about 7000 fairly conventional pre-coordinate type subject headings in thirteen broad subject categories. A hierarchical classification ("tree structure") of these terms is also available to the indexers and the search analysts. In January 1966, subheadings were introduced into the system. Subheadings, of which 53 were in use in 1966, are general concept terms (e.g., BIOSYNTHESIS, COMPLICATIONS) which can be affixed to main subject headings, thus effecting greater specificity through additional pre-coordination. Each subheading can only be used with main subject headings from specified MeSH categories. For example, the subheading ABNORMALITIES can only be used with Category A (anatomical) terms, while CONGENITAL is only applicable to Category C (disease) terms. These and other indexing conventions are spelled out in detail in a MEDLARS Indexing Manual revised annually. Appendix 1 of this report contains a sample page from MeSH, from the hierarchical (tree) display of MeSH terms, and the list of subheadings in use in 1967.

A demand search is presently conducted, on a Honeywell 800 computer, by serial search of the index term profiles of the 700,000 citations on magnetic tape. This search is essentially a matching process: the index term profiles of journal articles are matched against a search formulation, which is a translation of a subject request into the controlled vocabulary of the system. Requests for demand searches are mostly received by mail at NLM, either embodied in a letter or on a "demand search request form" (a specimen appears in Appendix 1); a higher proportion of the requests processed by regional MEDLARS centers are made by personal visit t the center. The search formulations are prepared, by search analysts, in the form of Boolean combinations (logical sums, logical products, and negations) of main subject headings and subheadings. A generic search (known at NLM as an "explosion") can be conducted by means of the tree structure. An "explosion on A9.44.44" means that a search is conducted on the generic term RETINA (identified as A9.44.44 in the tree structure) and all the terms subordinate to it in the tree structure, namely FUNDUS OCULI, MACULA LUTEA, and RODS AND CONES.

A search formulation may be constructed as a three-level strategy, which will result in a three-section printout (sections 4, 5 and 6) on the high-speed printer. Level 4 represents the broadest strategy employed by the search analyst. Level 5 introduces an additional restriction to this strategy, and produces a subset of the citations retrieved by the broader strategy. Level 6 introduces a further restriction and produces a subset of the citations retrieved by Level 5. For example, suppose the broadest strategy (Level 4) demands the retrieval of citations whose index term profiles match the following Boolean statement:

TERM A		TERM L
or	and	or
TERM B		TERM M

Level 5 might ask for the separation, from the citations retrieved by the strategy above, of those that had been indexed under TERM B and under TERM M (i.e., a subset of 4 is produced). Level 6 is more specific still, and requests that, of the citations matching the requirements of 5, any indexed under the term X are to be sorted out and printed separately. Note that it is possible to employ, for sorting purposes, in Level 5 and Level 6, an index term not forming part of the original (Level 4) searching strategy.

In the printout of the demand search bibliography, which is the normal product of a MEDLARS search, the citations are printed in the order: Section 6 (i.e., citations matching the requirements of Level 6), Section 5 (those citations matching the requirements of Level 5 that were not already printed in Section 6), Section 4 (those citations matching the general strategy that were not already printed in Section 5 or Section 6). This can be clarified by returning to the sample formulation mentioned above. Suppose that 205 citations satisfy the requirements of the general strategy

TERM A		TERM L
or	and	or
TERM B		TERM M

The profiles of 80 of these citations match the more stringent requirement of 5 (i.e., each citation is indexed under the term B and also under the term M). Of these 80 citations, ten have been indexed under the term X, and thus satisfy the most specific search requirement (Level 6). When the search is printed, these ten citations ("section 6" of the bibliography) appear first, followed by the 70 citations of section 5 (the 80 satisfying the Level 5 search requirement less the ten already printed in section 6), and finally the residue of retrieved citations is printed in section 4 (125 citations).

This three-level search capability is used in two ways within MEDLARS:

1. To produce a search of varying specificity in relation to the request. For example, assuming a request for literature on drug X, used to treat disease Y, particularly where this is shown to lead to side-effect Z, section 6 of the search printout may be designated to include citations relating specifically to the side-effect, while sections 5 and 4 relate more generally to the effects of drug X on disease Y.

2. Merely as a sorting device. For example, consider a request for toxins A, B, C, D, E and F. For convenience to the user, the searcher specifies that citations relating to toxin F be printed in section 6, citations to toxin E in section 5, and section 4 will cover "all other toxins", namely A, B, C, and D. Obviously, in this case the citations in section 6 are not more specific in relation to the request than those in section 5 or section 4. *

This 6-5-4 breakdown has been discussed in some detail because

a. it is somewhat peculiar to MEDLARS,

b. it tends to be confusing to people outside of NLM, and

c. an understanding of it is a prerequisite to the comprehension of certain of the results presented in Part 2 of this report.

The final product of a MEDLARS search is a computer-printed <u>demand search bibliography</u>, in up to three sections as discussed above, the citations usually appearing in alphabetical order by author within each section. Accompanying each bibliographic citation is a complete set of tracings (i.e., a record of all the index terms assigned to the article). A specimen page from such a bibliography is included in Appendix 1. So also is a sample search formulation.

* It is estimated that a little more than half the searches using the three-level sorting mechanism are of the first type.

OBJECTIVES OF THE TEST PROGRAM

The principal objectives of the test program may be summarized as follows:

1. To study the demand search requirements of MEDLARS users.

2. To determine how effectively and efficiently the present MEDLARS service is meeting these requirements.

3. To recognize factors adversely affecting the performance of MEDLARS.

4. To disclose ways in which the requirements of MEDLARS users may be satisfied more efficiently and/or more economically. In particular, to suggest means whereby new generations of equipment and programs may be used most effectively in satisfaction of demand search requirements.

In addition, the test was expected to produce further valuable benefits:

5. On the basis of test results, and analyses of failures, it would aid in establishing methods that could be used to implement a continuous "quality control" program for the MEDLARS operation.

6. The test would provide a corpus (of documents, requests, indexing, search formulations, and "relevance" assessments) that could be used for further tests and experimentation.

7. It would identify specialized areas that might require further experimentation and evaluation.

Test requirements

We assume that the prime requirements of demand search users relate to the following factors:

1. The coverage of MEDLARS (i.e., the proportion of the useful literature on a particular topic, within the time limits imposed, that is indexed into the system).

2. Its recall power (i.e., its ability to retrieve "relevant" documents, which, within the context of this evaluation, means documents of value in relation to an information need that prompted a request to MEDLARS).

3. Its precision power (i.e., its ability to hold back "nonrelevant" documents).

4. The response time of the system, (i.e., the time elapsing between receipt of a request at a MEDLARS center and delivery to the user of a printed bibliography).

5. The format in which search results are presented.

6. The amount of effort the user must personally expend in order to achieve a satisfactory response from the system.[2]

It follows, therefore, that the test had to establish user requirements and tolerances in relation to these various factors.

In particular, the test was designed to answer certain specific questions relating to the operating efficiency of the MEDLARS demand search service. These questions are enumerated below:

Overall performance

a. What is the overall performance level of MEDLARS in relation to user requirements? Are there significant differences for various types of request and in various broad subject areas?

Coverage and processing

a. How sound are present policies regarding indexing coverage?

b. Is the delay between the receipt of a journal and its a appearance in the indexing system significantly affecting performance?

Indexing

a. Are there significant variations in inter-indexer performance?

b. How far is this related to experience in indexing and to degree of "revising"?

c. Do the indexers recognize the specific concepts that are of interest to various user groups?

d. What is the effect of present policies relating to exhaustivity of indexing? In particular, is there a significant difference between retrieval performance for articles from "depth-indexed" and "non-depth-indexed" journals? What would be the effect of searching on only Index Medicus headings?

Index language

a. Are the terms sufficiently specific?

b. Are variations in specificity of terms in different areas significantly affecting performance?

c. Are pre-coordinate type terms and subheadings, which have been included to meet the requirements of <u>Index</u> <u>Medicus</u>, hindering the efficiency of retrieval by MEDLARS?

d. Is the need for additional precision devices, such as weighting, role indicators, or a form of interlocking, indicated?

e. Is the quality of term association in <u>MeSH</u> satisfactory?

f. Is the present "entry vocabulary" adequate?

<u>Searching</u>

a. What are the requirements of the users regarding recall and precision?

b. Can search strategies be devised to meet requirements for high recall or high precision?

c. How effectively can NLM searchers screen output? What effect does screening have on recall and precision figures?

d. What are the most promising modes of user/system interaction?

(1) Having more liaison with information staff at the local level?

(2) Having more liaison directly with MEDLARS search analysts?

(3) Certain alternative modes of interaction (e.g., user examination of proposed search strategy, or iterative search) not presently used in the MEDLARS operation?

e. What is the effect on response time of these various modes of interaction?

f. Are there significant differences in performance between the various MEDLARS centers?

<u>Input and computer processing</u>

a. Do input and data processing procedures, including various clerical functions, result in a significant number of search failures?

TEST DESIGN: GENERAL CONSIDERATIONS

From the point of view of the test design, the most critical problems faced were:

1. Ensuring that the body of test requests was, as far as possible, representative of the complete spectrum of "kinds" of requests processed.

2. Establishing methods for determining recall and precision figures.

Selection of user groups to participate in the evaluation

The sheer administrative problem of dealing individually, in various ways, with possibly several hundred individuals, and the volume of correspondence and other paperwork involved, made it impractical to take test requests completely at random as they were made to the system. Instead, a stratified sample was employed. The evaluation was based upon requests coming from a manageable number of organizations that agreed in advance to cooperate in the evaluation program. In this way, much of the direct liaison with the end users was carried out at the local group level, in particular by the librarians or information specialists of the organizations concerned.

A large part of the effort going into the test design was devoted to the identification of a number of user groups that would collectively form a suitable "test group" for the purpose of the evaluation program. The composition of the test group had to be based upon the following considerations:

1. Volume of requests. Based on past performance, the group must be likely to put a certain minimum number of requests in a restricted time period (say, 400 requests in 9 - 12 months).

2. Type of request. The "types" of requests to be expected from the test group must be representative of all the principal "types" of requests made to MEDLARS by the entire user population.

3. Type of organization. The test group must include representatives of the principal types of organization (e.g., research, clinical, development, regulatory) using the MEDLARS demand search service, in case there should be a significant difference in the ability of MEDLARS to satisfy their varying needs.

4. The composition of the group must be such that it allowed observation of the effects of the principal modes of user/system interaction operating in the system, namely:

1. Personal interaction: the requester comes directly to a MEDLARS center and negotiates his requirement directly with a search analyst.

2. No interaction: the request comes to a MEDLARS center by mail directly from the requester.

3. Local interaction: the request comes by mail, but through a local librarian or information specialist who may do something to modify it (e.g., by interviewing the requester or by conducting a preliminary literature search) at the local level.

A detailed study was carried out on the "search log books" recording demand searches completed by the National Library of Medicine in 1965. Based on expected volume of real-life requests, kind of organization, subject categorization of requests, and probable modes of user/system interaction, the following 21 user groups were finally selected as the "test user group" to participate in the evaluation program:

Harvard University (School of Medicine & School of Public Health)	ACADEMIC
UCLA	
Georgetown University	
Johns Hopkins University	
Albert Einstein College of Medicine	
University of Colorado	
University of Virginia	
National Institute of Neurological Diseases & Blindness	RESEARCH
National Cancer Institute	
Armed Forces Institute of Pathology	
Naval Medical Research Institute	
U.S. Air Force, School of Aerospace Medicine, Brooks AFB	
Smith, Kline & French Laboratories	PHARMACEUTICAL
Warner-Lambert Research Institute	
Boston City Hospital	CLINICAL
VA Hospital, District of Columbia	
VA Hospital, Pittsburgh	
Naval Medical Center	
Private practitioners *	
Food and Drug Administration	FEDERAL REGULATORY
National Communicable Disease Center	

* We decided to attempt to obtain the participation of some of the private practitioners, writing from their home or office, during the period of the test. This would add an additional user group that would be primarily clinical and it would allow us to observe (a) whether the requests from private practitioners were significantly different from other requests, and (b) whether MEDLARS could serve the needs of this group adequately.

This test group gives representation of all the major types of organization making use of MEDLARS, and it was expected, based on past performance, to submit a minimum of 400 requests in the twelve-month period assigned to the processing phase of the project. Moreover, the breakdown of 607 requests from these organizations into broad subject categories (see Table 1) satisfactorily resembled the subject-area breakdown of a larger group of 1136 requests from 105 centers selected from the 1965 search logs. The subject categories were selected and defined on the basis of the subject categories into

Table 1

Category breakdown of 607 requests from 21 user groups selected to participate in the study

Behavioral Sciences	35	5.5 %
Disease	206	34.6 %
Drug/Biology	70	11.4 %
Public Health	21	3.4 %
Preclinical Sciences	112	18.3 %
Drug/Disease	18	2.9 %
Technics	86	14.0 %
Drug and Chemical	47	7.7 %
Physics/Biology	18	2.9 %
	613	100.7 %

607 requests fell into 613 categories.

which Medical Subject Headings are grouped, as follows:

PRECLINICAL SCIENCES: Anatomy, biochemistry, cytology, genetics, immunology in general, microbiology, physiology, endocrinology, metabolism, nutrition, bacteriology, embryology.

DISEASE, INJURY AND PHYSICAL ABNORMALITY: Pathology. Nature and cause of disease and physical abnormalities, including experimentally induced disease. Symptoms. Natural course of disease. Includes biochemical aspects of disease (e.g., metabolic effects and histochemistry of diseased organs). Includes immunological studies on specific diseases, but not general studies on immunological properties (included under PRECLINICAL SCIENCES). Includes statistical and epidemiological requests. Excludes all human intervention (TECHNICS).

TECHNICS AND EQUIPMENT: Technics of diagnosis, treatment, measurement, analysis, and equipment used. Excludes drug therapy. Includes effects of technics.

DRUGS AND CHEMICALS:* All general studies on chemicals and drugs, excluding studies specifically on their effects. Excludes naturally-occurring body chemicals, but includes extracted and synthesized hormones, vitamins, etc.

BEHAVIORAL SCIENCES: Emotional and mental processes, including treatment, but excluding drug therapy and side effects.

PUBLIC HEALTH: Health of the community: hospitals, nursing, medical ethics, legal aspects, and all other studies in the social sciences and humanities relating to health of the community. Excludes epidemiology and statistics on disease.

DRUGS AND CHEMICALS/BIOLOGY (pharmacology and psychopharmacology): Effects of drugs and chemicals on the body, excluding deliberate use in treatment or diagnosis. Includes effects on behavior. Includes side effects.

DRUGS AND CHEMICALS/DISEASE AND DIAGNOSIS: Drug therapy and prophylaxis, including immunization.

PHYSICS/BIOLOGY: Effects of physical phenomena on the body.

*During the conduct of the evaluation program it was recognized that this category does not really exist as a separate entity. Requests to MEDLARS in this general area, although they appear more general on the surface, relate in some way to biological effects. The category was later dropped, all drug and chemical requests being put either in DRUG/BIOLOGY or DRUG/DISEASE.

It must be stressed here that this categorization does not represent an attempt to arrive at an authoritative classification of subject requests in the biomedical field. It is an empirically-derived classification based entirely upon the way that MEDLARS requests seemed, at least to one observer, to group themselves fairly naturally. We are satisfied that, for the purpose of ensuring that the "test-requests" were fully representative of the various "kinds" of requests being made to MEDLARS by the entire user population, this is a valid and useful categorization. The categorization is partly a conventional subject classification and partly a "viewpoint" or "method of approach" categorization. It cuts completely across certain conventional medical disciplines. For example it was found that 42 searches relating to dentistry could be categorized as follows: 14 fell into the area of PRECLINICAL SCIENCES, 11 fell under TECHNICS, 10 under DISEASES, six under DRUG/BIOLOGY, two under PUBLIC HEALTH and one under BEHAVIORAL SCIENCE.

A return rate (of relevance assessments) of about 75% was anticipated for the test searches, and it was felt that the approximately 300 searches that would thus be fully completed would be adequate to allow a meaningful performance breakdown by processing center, subject field, originating organization, and mode of interaction.

The 20 formal groups were invited to participate in the evaluation program by the Director of the National Library of Medicine, and all agreed to do so. Subsequent liaison was conducted between the author and the library or information staff of the organizations concerned.

Establishing the performance figures

The operating efficiency of MEDLARS was evaluated on the basis of its performance in relation to a number of demand search requests made, in a 12-month period, by individual physicians and other scientists affiliated with the twenty major medical organizations agreeing to cooperate in the study. It must be stressed here that, while the organizations comprising the test user group had agreed to cooperate in the evaluation program (e.g., the dean of a medical school or the director of a research institute agreed to the participation of the organization, and his librarian also promised assistance), the individual requesters knew nothing of the evaluation program until they submitted their requests. At that time they were asked to cooperate by allowing us to use their requests as "test requests". There is, then, no artificiality about the body of test requests. Each quite definitely represents an actual information need. For each of the test requests, a search was conducted and a computer printout of citations (demand search bibliography), which is the normal product of a MEDLARS search, was delivered to the requester. A duplicate copy of this printout was used in the extraction of a random sample of 25-30 of the retrieved citations. Photocopies of these sample articles were submitted to the requester for assessment, a second copy of each article being retained for analysis purposes. This figure of 25-30 represents an upper bound on the number of articles for which we felt we could reasonably expect to obtain careful assessments. If the search retrieved a total of 30 articles or less, we normally submitted all for assessment.

We believe categorically that, within the environment of an operating retrieval system, where the performance of the entire system is being evaluated, a "relevant" document is nothing more nor less than a document of some value to the user in relation to the information need that prompted his request. In other words, in a real operating situation, a "relevance assessment" is a value judgement made on a retrieved document. We also believe that, to obtain valid precision figures and other data for analysis purposes, value judgments carefully made on a sample of a complete search output are of much greater value than less careful assessments made grossly on the complete output.

A copy of the Form for Document Evaluation, which was attached to each article submitted for assessment, is shown in Figure 1. This form ascertained whether or not the requester was previously aware of the retrieved item, and asked him to assess the article as of major, minor or no value in relation to the information need that prompted his request to MEDLARS. Most importantly, the requester was required to substantiate these judgments by indicating why particular items are of major value, others minor, and yet others of no value. These substantiations are of great utility in the analysis of search results. To get some idea of the serendipity value of searches, the requester was asked to indicate whether or not an article, judged of no value in relation to the need that prompted his request, was in fact of interest in relation to some other need or project. Finally, if the user was unable to assess the article because of inability to read the language (approximately 45% of the material in the data base is in languages other than English), the form determined whether or not he intended to obtain a complete or partial translation of its contents.

While precision figures for a MEDLARS search present no particular problem, it is extremely difficult to estimate the recall ratio for a "real-life" search in a file of half a million citations. The only way to obtain a true recall figure is to have the requester examine, and make assessments on, each and and every document in the file. While this is feasible in certain experimental situations, it is obviously out of the question for a collection of the MEDLARS size. The size of the base also rules out any hope of obtaining recall figures by conventional random sampling among the documents not retrieved by a particular search.

We therefore estimated the MEDLARS recall figure on the basis of retrieval performance in relation to a number of documents, judged relevant by the requester, but found by means outside MEDLARS. These documents could be, for example,

1. documents known to the requester at the time of his request,

2. documents found by his local librarian in non-NLM generated tools,

Figure 1

BoB No. 68-R-938
App. Exp. 12/31/67

NATIONAL LIBRARY OF MEDICINE
Bethesda, Maryland

Request No. ______
Document No. _____

MEDLARS EVALUATION PROJECT

Form For Document Evaluation

1. Were you previously aware of the existence of this article?

   Yes [ ] How did you learn of its existence?

   No [ ]

2. By checking the appropriate box, please evaluate this article in relation to the information need that prompted your request to MEDLARS.

   (a) Of major value to me in relation to my information need [ ]
   Please explain why:

   (b) Of minor value to me in relation to my information need [ ]
   Please explain why:

   (c) Of no value to me in relation to my information need [ ]
   Please explain why:

   Were you glad to learn of its existence because of some other need or project:

   Yes [ ] Please explain why:

   No [ ]

   (d) Unable to make an assessment because of language of the document [ ]

   Do you intend to take any steps to determine the contents of this foreign language document?

   Yes [ ] Please specify what steps:

   No [ ] Please explain why:

FIGURE 2

BoB No. 68-R-938
App. Exp. 12/31/67

NATIONAL LIBRARY OF MEDICINE
Bethesda, Maryland

MEDLARS EVALUATION PROJECT

Record of Known Relevant Documents

Name of Requester K. Nagarajan & R.L. Beaudoin

Organization NMRI, NNMC, Bethesda

Search No.________

Instructions: Please list all papers published since July 1963 already known by you to be relevant to the subject of your request to MEDLARS. Check the **appropriate column to indicate whether they are of major or minor value** in relation to the information need that prompted your request. If they were **found as a** direct result **of a literature search in** Index Medicus, **please check the last column.**

Articles	Major Value	Minor Value	Index Medicus
1. Effect of the antimalarials Chloroquine on the Phospholipid metabolism of Avian Malaria and heart tissue Amer. Journ Trop. Med. Hyg. (1966) <u>15</u>, 818-822.	x		
2. The incorporation of radioactivity from C^{14} Glucose into the soluble metabolic intermediatesof malrial parasites Amer. Journ Trop. Med. Hyg. (1966) <u>13</u>, 515-524.		x	
3.			
4.			
5.			
6.			
7.			
8.			
9.			

3. documents found by NLM in non-NLM-generated tools,

4. documents found by some other information center, or

5. documents known by authors of papers referred to by the requester.

For every test request we attempted to obtain a record of any articles, within the time span of MEDLARS, that the requester already knew to be relevant to the subject of his request. An example of a completed *Record of Known Relevant Documents* is included as Figure 2. This form was completed by the requester after he had submitted his request but before he received the results of a MEDLARS search.

If the requester was able to supply a substantial quantity of citations not found by him in *Index Medicus* (citations found through direct search of *Index Medicus* should theoretically introduce a substantial bias into the recall estimate, since MEDLARS indexing is *Index Medicus* indexing plus), this was accepted as the recall base without further expansion. However, if the requester knew of no articles, or only one or two, an attempt was made to find additional potentially relevant items *by means outside of the system*. These might be articles found by the librarian of the organization submitting the request, searching in tools not generated by the National Library of Medicine. Alternatively, they could be found by conventional manual literature searches conducted by members of the Evaluation Group in non-NLM generated tools held at the Library. In some cases, the one or two citations supplied by the requester would yield additional possibly relevant items, by means of a search in the *Science Citation Index*, or through direct contact with the authors of these known relevant papers. Occasionally it was possible to obtain additional items from a specialized information center such as the Parkinson's Disease Information and Research Center at Columbia University.

Although all of these methods of augmenting the recall base were tried in the current evaluation, experience showed that conventional manual searching at NLM was the method most likely to expand the recall base with the minimum of effort. The documents found by these various methods, extraneous to MEDLARS, were considered no more than "possibly relevant". They were not incorporated into the *recall base* until the requester had examined them and judged them as of some value in relation to his information need. To achieve this, these additional items were interspersed with the *precision set* (i.e., the articles selected by random sampling from the MEDLARS search printout). The requester then assessed the enlarged set at one time.

Table 2 illustrates the way in which this method of obtaining a recall estimate works. In this instance, the requester is able to name 2 relevant documents and his local librarian finds an additional 7 which she believes to be relevant to the physician's request. The user, asked to make assessments of these 7 documents, judges 4 to be relevant. We now have 6 known relevant documents upon which to base our recall figure. If all are in

TABLE 2

	Documents Found Outside of MEDLARS	Documents Judged Relevant
REQUESTER	2	2
LOCAL LIBRARIAN	7	4
NLM STAFF		
OTHER CENTER		
AUTHORS OF PAPERS REFERRED TO BY REQUESTER		
Totals	9	6

MEDLARS RETRIEVES 4/6 RECALL RATIO FOR SINGLE REQUEST 4/6 x 100 = 66%

the MEDLARS data base, but only 4 are retrieved, we can say that the recall ratio for this search is 66%. This method works equally well, of course, whether the "possibly relevant" documents are discovered by the local librarian, NLM staff (in non-NLM tools), or by some other specialized information center, or are named by the author of a relevant paper referred to by the original requester.

Another way of considering this method of obtaining a recall estimate is illustrated by Figure 3.

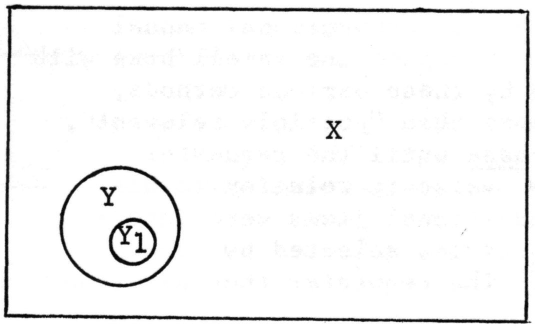


Figure 3

The area X represents the entire MEDLARS collection of half a million items. For any particular request made to the system, if the requester examined each and every item in the collection, he would be able to identify a subset, Y, of items which he considered of value in relation to his information need. All other items in the collection (X-Y) are of no value (i.e., "not relevant"). Unfortunately, except by complete examination of the collection there is no foolproof method of establishing for any one request the exact subset Y of relevant items. However, we can establish a subset of the subset. That is, by methods outlined above, we can find some group, Y_1, of articles which the requester agrees to be relevant. We now establish the recall estimate on the basis of the performance of the system in relation to this particular group of relevant items. Thus, if we know ten relevant articles within the data base, and MEDLARS retrieves seven of these, but misses three, we say that the MEDLARS recall ratio for this search is 70%, the assumption being that the "hit rate" for the group of documents Y_1 will approximate to the hit rate for the larger group Y.

It must be remembered that recall and precision figures are merely yardsticks by which we measure the effect of making certain changes in our system or in ways of operating the system. Although the recall estimate obtained by the present methodology may be slightly inflated or slightly deflated in relation to "true recall", since the method used to obtain the estimate was held constant throughout the evaluation program, the figures are still valid indicators of performance differences in various situations. The use to which these figures were put is discussed in detail in Part 2 of this report.

THE PRETEST

A pretest was conducted with 20 demand search requests made to MEDLARS in the period January-March 1966. The pretest was intended primarily to:

1. Simulate the modus operandi proposed for the main test program.

2. Test the proposed forms.

3. Obtain some preliminary figures for the general performance range of MEDLARS, and

4. Test certain hypotheses upon which the test design was founded (for example, the ability of requesters to name some relevant documents before the MEDLARS search).

The pretest proved adequate as a simulation of the main test program, and forms and procedures were usefully modified as a result of experience gained in the pretest. In the pretest we were able to obtain an average of five "known relevant documents" per requester. The MEDLARS recall estimate, averaged over the 20 requests, was 62% while the average precision ratio was 59.2%.

PROCEDURES USED IN THE CONDUCT OF THE TEST

Between August 1966 and July 1967 some 410 test requests from 21 user groups were processed by the National Library of Medicine and by the MEDLARS centers at the University of Colorado, Harvard, the National Institutes of Health, and the University of California at Los Angeles. At first all requests (at least where the requester indicated willingness to cooperate-- and over 90% were willing) from the 20 formal groups were accepted as test requests. Later, however, when we felt that we had processed sufficient requests from any one particular user group, no further requests from this group were treated as "test requests". This was done in an attempt to avoid collecting a disproportionate number of requests from any one organization. As it was, we received an unexpectedly large number of requests from Harvard University, and this organization was the first to be cut off from the test processing. On the other hand, certain organizations (for example, the Veterans Administration Hospital in Pittsburgh) submitted fewer requests in the test period than we expected based on the 1965 statistics. From the beginning, it proved very difficult to include in the program requests from private practitioners. A very small proportion of the MEDLARS requests are submitted by this group, those that are are difficult to identify as coming from the true private practitioner (as opposed to a specialist affiliated with some university, but happening to write to MEDLARS from his home or office), and it was usually difficult to persuade them to cooperate in the study. For this reason, we were only able to obtain six completed test searches from private practitioners.

It is worthwhile devoting some time to a more detailed description of how exactly the test requests were processed. They arrived at a MEDLARS center in one of three ways:

a. By personal visit of a requester to a MEDLARS center and negotiation of the request directly with a search analyst. This was true of all the requests emanating from requesters at the University of Colorado and UCLA, and the great majority of requests made by the staff of the National Institutes of Health and Harvard University. These organizations themselves operate MEDLARS processing centers.

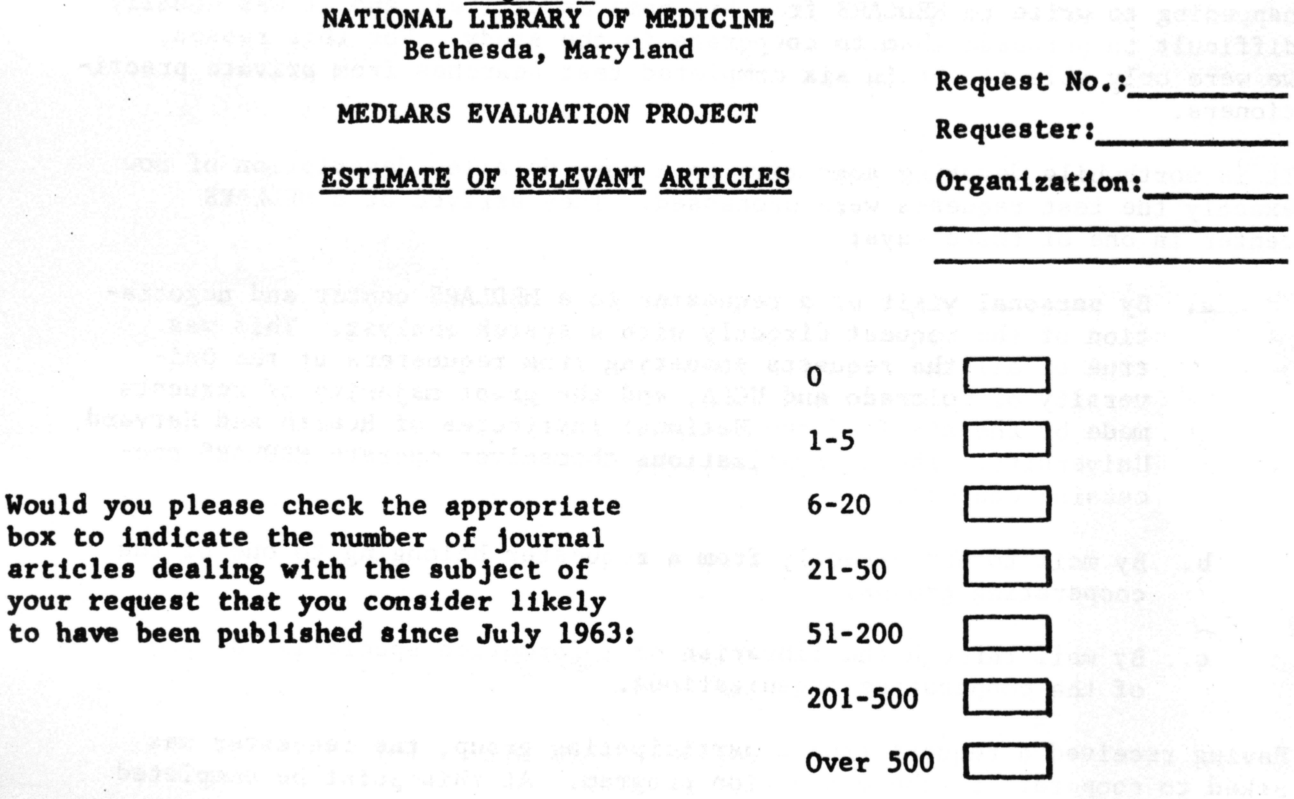
b. By mail to NLM directly from a requester belonging to one of the cooperating groups.

c. By mail through the librarian or information specialist at one of the cooperating organizations.

Having received a request from a participating group, the requester was asked to cooperate in the evaluation program. At this point he completed two forms, the Record of Known Relevant Documents (Figure 1) and the Estimate of Relevant Articles (Figure 4). Cooperation was secured by the search analyst, by the local librarian, or directly by the Evaluator, depending upon how the request was received. The two test forms, together with a xerox copy of the request statement, were delivered to the Evaluator, thus allowing the request to be logged in and numbered as a "test request". The request was then formulated and searched in the normal way, with the exception that it was labeled as a "test request" to ensure that the Evaluator received the further records needed to conduct the study. The forms specially collected for purposes of the test were not available to the search analysts preparing formulations for these test requests. A test search having been completed, the demand search bibliography was forwarded to the requester as usual, a second copy of this, together with a copy of the search formulation, being submitted to the Evaluator.

The evaluation copy of the search printout was used to extract a random sample of retrieved citations. A random number table was used to provide a "random start". Thereafter, a regular sampling interval was adopted, thus allowing the three separate segments of the search (6, 5 and 4), where the search was so divided, to be correctly sampled in proportion to their size.

Figure 4



NATIONAL LIBRARY OF MEDICINE
Bethesda, Maryland

MEDLARS EVALUATION PROJECT

ESTIMATE OF RELEVANT ARTICLES

Request No.:__________

Requester:__________

Organization:__________

____

____

Would you please check the appropriate box to indicate the number of journal articles dealing with the subject of your request that you consider likely to have been published since July 1963:

0 □

1-5 □

6-20 □

21-50 □

51-200 □

201-500 □

Over 500 □

The sample of citations was delivered to the Reference Services Division of the National Library of Medicine, and two complete xerox copies of each article were provided. No attempt was made to wait for journal parts in use or at binding. It was for this reason that slightly more than 25 citations were selected by sampling from the search printout, so that the eventual set delivered to the requester would be 25 or close to that number. Where the complete search retrieved 30, or fewer, citations we would normally photocopy all the articles involved or at least all those on the shelf at the time. From examination of the results in Part 2 of this report, it can be seen that the number of articles actually assessed varies from search to search depending upon: (a) the number of articles retrieved, (b) the number selected by random sampling, (c) the number actually on the shelf when requested, and (d) the number actually assessed by the requester (some could not be assessed because the requester could not read the language of the article, and in one or two cases the requester failed to return all of the evaluation forms).

One complete set of the articles forming the random sample was set aside for submission to the requester, the second set being filed for analysis purposes. All articles found by parallel manual search, and thus forming part of the recall base for the search, were also photocopied in duplicate. These recall base articles, for which evaluations were required, were interspersed among the articles forming the random sample (precision set), except that unwanted duplicates(of articles happening to fall both in the recall base and the precision set) were discarded. Each article in the requester's set was given a unique number consisting of the search number and the item number (1/1 was the first item in the sample for search #1) and these numbers were transferred to the evaluator's set of articles and to the Form for Document Evaluation (Figure 1) attached to the front of each article in the requester's set. In addition, the Evaluator's copy of those articles falling into only the recall base for the search were marked "recall only", while those falling into both recall and precision bases (i.e., articles found by parallel manual search and also happening to fall in the random sample selected from the search printout) were marked "recall and precision". The requester's set of photocopies, each with a Form for Document Evaluation stapled to it, was now mailed to him, together with a covering letter, a set of Notes on Form for Document Evaluation, and two additional brief forms, one asking about the timeliness of the MEDLARS service and the other inviting him to rephrase his request should he feel that the search results indicated that his original request statement was inadequate. A sample of each of these additional enclosures in included in Appendix 2.

Of course, the requester was allowed to keep the photocopies for his own files. He was merely required to return the completed forms. When these arrived at NLM, each Form for Document Evaluation was attached to the article to which it related in the file set. This search was now ready for analysis.

Preliminary analysis consisted of two parts:

1. Derivation of performance figures for the search.

2. Analysis of reasons for search failures.

Derivation of performance figures

Where necessary, the first thing to be done was to divide the file set into three parts: the "recall only" set, the "precision only" set, and the "recall and precision" set. The recall base articles were dealt with first. Each article in the "recall only" set was now checked against the search printout to determine whether or not it had been retrieved. The same thing was done for each item on the requester's completed Record of Known Relevant Documents (Figure 2). These articles or citations were now marked "retrieved" or "not retrieved" as appropriate. The articles in the "recall and precision" set did not require checking against the printout: obviously, since they fell in the random sample, they had been retrieved by the search. Each "not retrieved" item (among the "recall only" articles or the citations on the Record of Known Relevant Documents) was now checked against the author indexes of Index Medicus and Cumulated Index Medicus to ensure that it was in fact in the MEDLARS data base. An article not thus found was obviously excluded from the recall base of the search. It was now possible to derive a recall estimate for the search as illustrated in Table 2. The complete recall base for a search consists of (1) any articles listed on the Record of Known Relevant Documents and subsequently proved to be in the MEDLARS data base, and (2) any articles found by parallel manual search, judged relevant by the requester when submitted to him in photocopy form, and subsequently proved to be in the MEDLARS data base. The recall ratio is the proportion of this recall base set retrieved by the MEDLARS search. Returning to Table 2, in this example the requester listed two articles on his Record of Known Relevant Documents, and both are in the MEDLARS base. Parallel manual search turned up seven items and these were submitted to the requester for assessment along with the random sample (precision set). However, the requester judged only four of these to be relevant,* so that the full recall base consists of six articles. On checking the search printout it was found that four of these articles were retrieved, but two were missed. The two missed articles are in the data base, so the recall ratio for the search is 4/6 x 100, or 66.7%. A separate recall ratio was also calculated for the recall base articles judged of major value by the requester.

Having got the recall figures out of the way, the random sample of articles (i.e., the "precision only" set and the "recall and precision" set) was reconstructed and the relevance assessments (value judgments) tabulated as shown in Table 3. This allows the derivation of the precision ratio: the total of all articles judged of value over the total of articles assessed. In this case 18 articles were assessed: four judged of major value, six of minor value, and eight of no value. The requester looked at an additional five items (making the total random sample submitted to him 23 articles) but could not judge their value because they were

* The Evaluator's "personal precision ratio" during the study was about 80%. That is, approximately 80% of the articles found by parallel manual search were judged relevant.

Table 3

	MAJOR	MINOR	NO VALUE	NOT ASSESSED
KNOWN IN ADVANCE	3	1		
NOT KNOWN	1	5	8	5

written in a language with which he was unfamiliar. The precision ratio for this search is, then, 10/18, or 55.5%, while the proportion of major value articles retrieved is 4/18, or 22.2%. Since this is a true random sample among the retrieved citations, we can extrapolate confidently to the complete search. In other words, if the requester looked at all the articles cited in the demand search bibliography, he would judge approximately 55% to be of some value to him in relation to his information need, and approximately 22% of the articles retrieved will be of major value to him.*

Another ratio of some interest is the <u>novelty ratio</u>, which indicates what proportion of the articles judged of value by the requester was brought to his attention for the first time by the MEDLARS search. From the results of Table 3 we can derive the overall novelty ratio of 6/10, or 60% (i.e., six of the ten articles judged relevant were brought to the attention of the requester for the first time by the MEDLARS search, the other four being known to him prior to receiving the MEDLARS search results). We can also derive separate novelty ratios for major and minor value items.

The novelty ratio allows us to make certain inferences on the familiarity of various requesters with the literature of their subject field, and on the contribution of the MEDLARS searches to the satisfaction of disparate information needs. Certain requesters are quite familiar with the literature relating to their research topic. The MEDLARS search is conducted to insure that they have not overlooked articles of central importance, and to bring to their attention, for the first time, certain articles of peripheral interest. Other requesters, approaching a particular area for the first time, are unfamiliar with the literature and virtually all relevant items retrieved are new to them (i.e., the MEDLARS search has a high <u>novelty ratio</u>).

The final performance figures derived for a search were recall ratios and precision ratios for the separate sections, where the search had been so ordered, remembering, of course, that "4" equals the full search and thus includes both section 5 and section 6, and that section 5 includes section 6. When these results are tabulated they normally display the familiar inverse relationship between recall and precision, as the following specimen

* MEDLARS is used almost exclusively for comprehensive or semi-comprehensive searches, and not to discover "a few relevant items".

indicates:

	Recall ratio	Precision ratio
Full search (4)	10/14= 71.4%	11/23= 47.8%
Section 5 only	3/14= 21.4%	6/7= 85.7%
Section 6 only	1/14= 7.1%	2/2= 100%

Analysis of reasons for search failures

Having calculated and recorded the performance figures for a test search, the next step involved the detailed intellectual analysis of reasons why recall and precision failures occurred. Referring once more to the sample recall and precision results tabulated in Table 2 and Table 3, it can be seen that, in this particular search, we are faced with the analysis of:

a. two recall failures (two of the six "known relevant" articles were not retrieved), and

b. eight precision failures (eight of the 18 articles assessed by the requester were judged of no value).

It must be stressed here that the two recall failures and the eight precision failures are not the only failures occurring in the search. They are the only ones that we know of and as such they are accepted as exemplifying the complete recall failures and precision failures of the search (i.e., they a symptomatic of problems occurring in this search).

The "hindsight" analysis of a search failure is the most challenging aspect of the evaluation process. It involves, for each "failure", an examination of the following:

1. The full text of the document itself.

2. The indexing record for this document (i.e., the record of index terms assigned, which is obtained by printout from the magnetic tape record).

3. The request statement.

4. The search formulation upon which the search was conducted.

5. The requester's completed assessment forms, particularly the reasons for articles being judged "of no value", and any other information supplied by the requester (e.g., in covering letter, by telephone, or on the form recording his revised request statement).

On the basis of all of these records, a decision is made as to the prime cause or causes of the particular failure under review.

Almost all of the failures can be attributed to some aspect of indexing, searching, the index language (i.e., MeSH and its auxiliaries), computer processing, or the area of interaction between the requester and the system.

All of this intellectual analysis was conducted by the author within the present evaluation program. In other words, the author made decisions as to which specific aspect of the system was primarily responsible for the failure under review. Although, on the surface, this type of analysis would appear to be the purely subjective decision of a single individual, in the MEDLARS evaluation it was not so. The attribution of system failures was, in a sense, the joint decision of the requester and the Evaluator because the requester's statement of why a particular document was "of no value" was often a good guide to where, in fact, the system had failed. This will become evident in the presentation of the results in Part 2 of this report. Wherever possible, for any one failure, a single "most critical" cause was isolated. In some instances, however, it was not possible to identify a single cause because two functions of the system were equally concerned. For example, for certain recall failures we can say that the article would have been retrieved if the indexer had used the additional term X. On the other hand, and equally important, had the searcher generalized from the adopted strategy A_1 and B and C to the reasonable approach of A and B and C, the article would also have been retrieved. In such cases, the failure was attributed jointly to indexing and searching, or whichever other elements of the system were jointly responsible.

While the ultimate decision as to the source of any failure was made by the author, he had the benefit of being able to consult with indexers, searchers, and vocabulary specialists on the staff of the Library, and did so in cases of problems. In certain other instances, he clarified various "doubtful" relevance assessments by contacting the requester. While the author does not claim to have made the only correct decision as to source of failure in all cases (nor does he expect 100% agreement with all decisions), he is satisfied that the decisions made have been generally consistent. He has been gratified to discover that his original decisions were usually replicated when it was found necessary to re-examine the data for certain special analyses.

A specimen of a complete search analysis, exactly as recorded by the Evaluator, is presented as Appendix 3. A complete set of analyses for the 302 searches, upon which the results of this study are based, is on file at the National Library of Medicine and available for consultation.