NATIONAL LIBRARY OF MEDICINE

EVALUATION of the MEDLARS DEMAND SEARCH SERVICE

January 1968

NATIONAL LIBRARY OF MEDICINE

EVALUATION of the MEDLARS DEMAND SEARCH SERVICE

January 1968

F.W. Lancaster
Deputy Chief, Bibliographic Services Division
National Library of Medicine

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

"IN ORDER TO SURVIVE, A SYSTEM MUST MONITOR ITSELF, EVALUATE ITS PERFORMANCE, AND UPGRADE IT WHEREVER POSSIBLE."

---

* From report GER 12760 Measures of Effectiveness and Criteria for Evaluation of a Document Processing System. Rome Air Development Center, 15 July 1966.

PREFACE

This report presents the results of a detailed analysis by the National Library of Medicine of the performance of MEDLARS (its Medical Literature Analysis and Retrieval System) in relation to 300 actual requests made to the system in 1966 and 1967. Thanks to this study, we now know considerably more about the requirements of MEDLARS users, how well the system is meeting these requirements, and what we must do to improve the overall performance level. The investigation is timely: the Library is now planning a second-generation MEDLARS, and the design of MEDLARS II should benefit greatly from our knowledge of factors affecting the performance of the present system.

Since this is the first large-scale evaluation of a major operating information system, and because of the diversity of subject areas covered by the study, it should be of considerable interest to the scientific community at large. Some readers, of course, may wish to take exception to parts of the methodology of the study or even view some of the analyses with reservation. In an effort to make the study as objective as possible, the design and results were reviewed by a distinguished outside advisory committee to whom we are most grateful.

To remain responsive to the demands of its users, a large scientific or technical information system must examine itself critically. We hope that a major benefit of this investigation will be the establishment of a program for the continuous quality control of MEDLARS products and services.

Martin M. Cummings, M.D.
Director
National Library of Medicine

CONTENTS

		Page
PART 1	DESIGN AND EXECUTION OF THE EVALUATION PROGRAM	
	Introduction	8
	MEDLARS: general background	9
	Objectives of the test program	12
	Test design: general considerations	15
	Selection of user groups	15
	Establishing the performance figures	19
	The pretest	26
	Procedures used in the conduct of the test	27
	Derivation of performance figures	30
	Analysis of reasons for search failures	32
PART 2	THE TEST RESULTS	
	Overall performance figures	35
	The individual ratios	38
	Average MEDLARS performance for the test requests	40
	Analysis of causes of recall and precision failures	46
	Analyses of failures: explanatory notes	46
	Recall and precision failures attributable to the indexing subsystem	51
	Failures due to exhaustive indexing or to lack of exhaustivity	54
	Effect of exhaustivity levels	58
	Failures due to lack of specificity in indexing	61
	General observations on exhaustivity and specificity of indexing within MEDLARS	62
	Evaluation of indexing as part of the input subsystem	64

Recall and precision failures attributable to the searching subsystem 65

Recall losses resulting from failure to cover all reasonable approaches to retrieval 65

Recall and precision failures due to variations in exhaustivity of the formulation 67

Recall and precision failures due to variations in specificity of the formulation 70

Use of "weighted" index terms 76

Other causes of searching failures 79

Joint causes of system failures 81

Effect of the 6-5-4 levels on recall and precision figures 81

Recall and precision failures attributable to the index language 82

Failures due to false coordinations and incorrect term relationships 91

General observations on the MEDLARS index language 98

The relationship between indexing, index language, and searching 100

Recall and precision failures attributable to the area of user-system interaction 101

Improving request statements 117

Experimentation with modes of interaction 125

Recall and precision failures attributable to computer processing 127

The novelty ratio 128

Factors affecting performance of a MEDLARS search 130

Variations in performance between five MEDLARS centers 145

MEDLARS indexing coverage 157

Foreign literature usage factors 159

Journal usage factors in MEDLARS 165

Effect of MEDLARS response time 174

The serendipity value of MEDLARS searches 175

Output screening 176

Indexer consistency 180

Requests rejected by MEDLARS 183

PART 3 CONCLUSIONS AND RECOMMENDATIONS

Overall MEDLARS performance 185

Upgrading the performance of MEDLARS 191

User-system interaction 193

The MEDLARS index language 193

The MEDLARS searching strategies 195

The MEDLARS indexing 197

Computer processing 199

The relationship between indexing, searching and MeSH 200

Use of foreign language material in MEDLARS 200

The search printout as a content indicator 201

Continuous quality control of the MEDLARS operation 201

Future use of the MEDLARS test corpus 202

PART 4 APPENDICES

Appendix 1. Samples of MEDLARS vocabulary, sample request and search formulation, sample printout 206

Appendix 2. Specimens of forms used in evaluation 213

Appendix 3. Specimen analysis worksheet 218

Appendix 4. Complete set of recall and precision ratios 230

Appendix 5. Analysis of the specificity of search formulations 249

Appendix 6. Analysis of the effect of subheadings on false coordinations and incorrect term relationships 264

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 demand search bibliography, 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?

307-006 O-68—2

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 Index Medicus, 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 MeSH satisfactory?

f. Is the present "entry vocabulary" adequate?

Searching

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?

Input and computer processing

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:

   - - 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 Search No._____

Organization NMRI, NNMC, Bethesda

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) 15, 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) 13, 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.

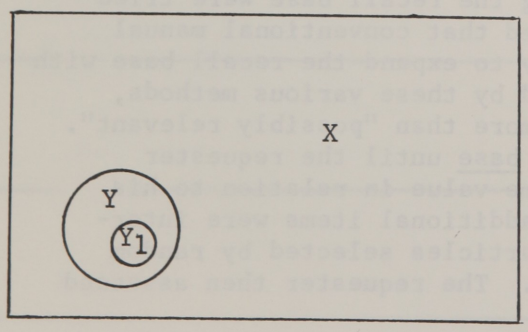


Figure 3

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 informa-

tion 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 novelty ratio, 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 novelty ratio).

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.

PART 2

THE TEST RESULTS

OVERALL PERFORMANCE FIGURES

In the period August 1966 - July 1967, 410 test requests were processed to the point of submitting photocopies of sample articles to the requesters. From these, 317 sets of relevance assessments had been returned as of October 15, 1967 (i.e., a 77% return rate). A total of 302 of these searches were completely analyzed, and these analyses form the basis of the results presented in this section of the report. The fifteen searches completed but not analyzed are:

1. Searches for which sets of assessments were received after the cut-off date established (October 1, 1967), and

2. Searches for which sets of assessments were received more than four months after the sample photocopies were submitted to the requester. It was arbitrarily decided that some such cutoff should be established to reduce the likelihood of including evaluations made so long after submission of the request that the user's orientation may have changed drastically.

The 302 searches finally analyzed were conducted for requesters affiliated with the 21 user groups according to the following distribution:

Harvard University	45
University of Colorado	28
Naval Hospital, National Naval Medical Center	25
Johns Hopkins University	21
National Cancer Institute	20
Veterans Administration Hospital, Washington, D. C.	20
Naval Medical Research Institute	19
University of Virginia	19
U.S.A.F. School of Aerospace Medicine	13
Albert Einstein College of Medicine	12
University of California at Los Angeles	11
Armed Forces Institute of Pathology	10
Georgetown University	10
National Institute of Neurological Diseases and Blindness	9
Veterans Administration Hospital, University Drive, Pittsburgh	8
National Communicable Disease Center	7

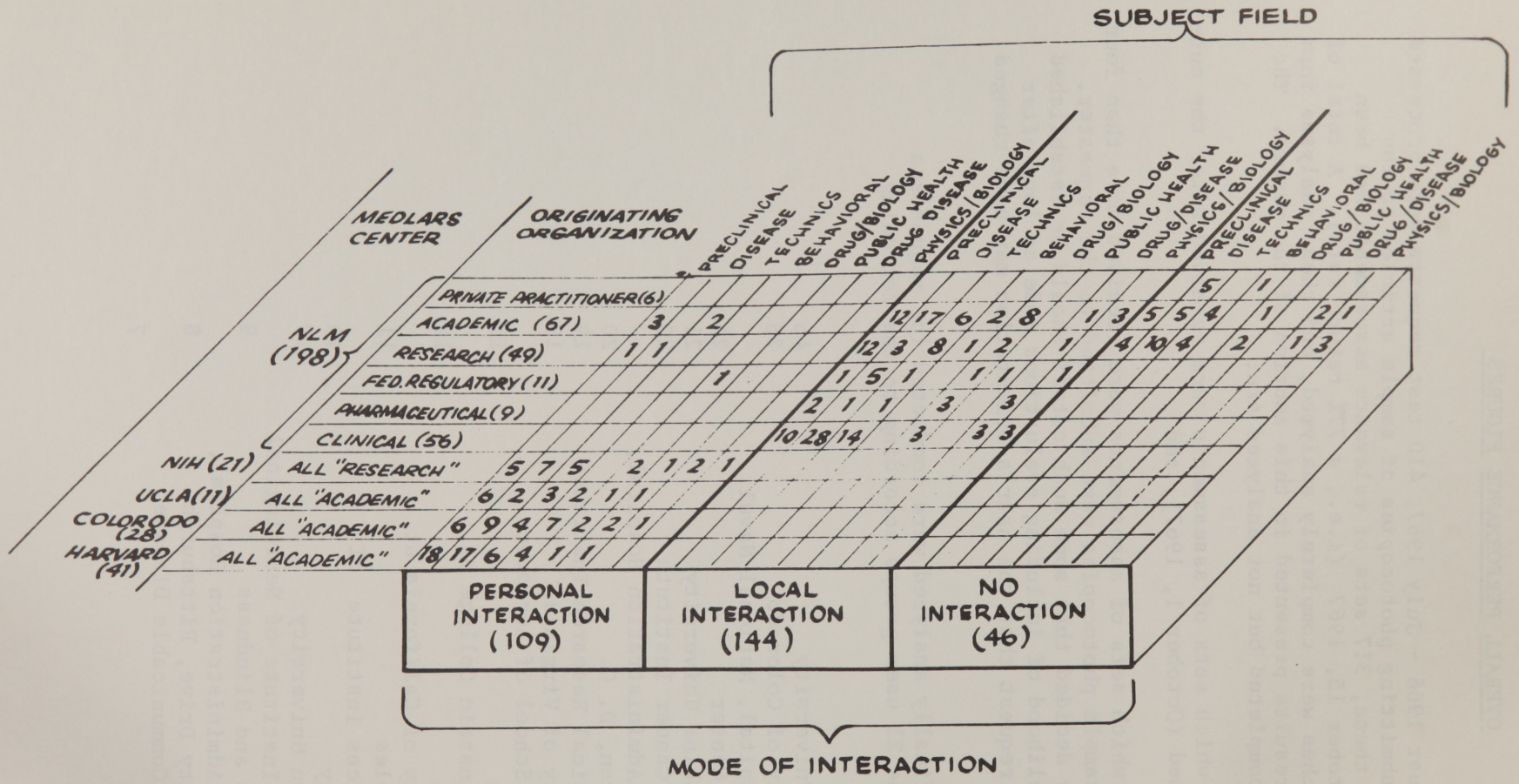

SUBJECT FIELD
MEDLARS CENTER
ORIGINATING ORGANIZATION
PRECLINICAL
DISEASE
TECHNICS
BEHAVIORAL
DRUG/BIOLOGY
PUBLIC HEALTH
DRUG/DISEASE
PHYSICS/BIOLOGY
NLM (198)
PRIVATE PRACTITIONER(6)
ACADEMIC (67)
RESEARCH (49)
FED. REGULATORY (11)
PHARMACEUTICAL (9)
CLINICAL (56)
NIH (21)
ALL "RESEARCH"
UCLA (11)
ALL "ACADEMIC"
COLORODO (28)
ALL "ACADEMIC"
HARVARD (41)
ALL "ACADEMIC"
PERSONAL INTERACTION (109)
LOCAL INTERACTION (144)
NO INTERACTION (46)
MODE OF INTERACTION

Figure 5

Smith, Kline & French Laboratories	7
Private Practitioners	6
Food and Drug Administration	5
Boston City Hospital	5
Warner Lambert Research Institute	2

Figure 5 shows how these 302 requests break down by mode of interaction, by "kind" of organization, by subject field of request, and by the MEDLARS center processing the search. The complete set of recall and precision ratios for the 302 searches analyzed is presented in Appendix 4. These performance figures require some explanation. For each search the following data are given:

1. The total number of citations "retrieved" by the search and delivered to the requester as a demand search bibliography on the subject of his request. In cases where this figure differs from the total of citations satisfying the search criteria (i.e., the citations actually "retrieved" by the computer) the latter figure is given parenthetically.

2. The precision ratio for the search, based upon the random sample of articles actually assessed by the requesters. For example, in search #1, the requester assessed 24 articles and judged 19 to be relevant. This gives an overall precision ratio of 19/24, or 79.2%. A precision ratio based on the major articles only is also presented. Thus, of the 24 articles assessed in the first search, six were judged to be of major value to the requester. Thus, the major value precision ratio for this search is 25%. In other words, if the requester examined all of the articles cited in the search printout, he would be likely to find 25% of major value to him in relation to the information need prompting his MEDLARS request.

3. The recall ratio for the search based upon the complete recall set (i.e., both the articles named by the requester in advance and those found by parallel manual search and subsequently judged relevant by the requester). Thus, in search #1 the full recall base of known relevant documents was 17. Of these, 15 were retrieved by MEDLARS and two were missed. The recall estimate for the search is thus 15/17, or 88.2%.

A separate recall ratio is given for the articles judged of major value by the requester. Seven of the recall base articles in search #1 were major value items, and five of these were retrieved. The major value recall ratio is, then, 5/7 or 71.4%.

4. The components of the recall ratio. The last four columns of the table present separate ratios for individual components of the recall base. For example, in #3 the complete recall base consists of 12 articles and the overall recall estimate is 11/12 (91.7%). Four of the 12 articles were cited by the requester before the MEDLARS search was conducted, and all of these were retrieved (recall ratio for this

component = 4/4, or 100%), while seven of the eight articles found by manual search, and judged relevant by the requester, were retrieved (recall ratio for this component = 7/8, or 87.5%). Note that these two components of the recall base are not always mutually exclusive, although they usually are. In other words, the "found manually" set may overlap the "known by requester" set. This reflects the fact that a parallel manual search was conducted independently of the requester's submissions, and may in fact have been conducted before the requester's Record of Known Relevant Documents was received. This can be seen in the results for search #4, where the only item known in advance by the requester was also found by the parallel manual search.

Finally, the "best set" recall ratio is presented. The requester, when completing his Record of Known Relevant Documents, was asked to indicate which items, if any, he found by means of direct search in Index Medicus. Theoretically, the inclusion of this group of documents in the recall base could lead to a substantial bias of the results in favor of MEDLARS. If the requester can find a particular article by direct search of Index Medicus under, presumably, the most likely subject headings, its retrieval should present no particular problem in the demand search module, which is based on the Index Medicus indexing plus additional terms used only for machine retrieval purposes.

The "best set", therefore, is the recall base with this group of articles omitted (i e., the base with the least possibility of bias). In other words, the "best set" comprises the articles found by parallel manual search, and judged relevant by the requester, and the articles named in advance by the requester but not found by him in Index Medicus. In actual fact, the "best set" differs very little from the overall recall base, because comparatively few of the articles cited by requesters were found by searching Index Medicus.

The individual ratios

For three of the 302 searches we were unable to obtain a recall base, but for the remaining 299 searches we have a precision ratio and a recall ratio. The precision ratio is based on the random sampling of the retrieved citations and we can have a certain amount of confidence in the figure quoted for each individual search. The recall ratio, on the other hand, must be taken as a recall estimate only. As discussed at some length in Part 1 of this report, we have no practical method of establishing "true recall" for a search in a file of over half a million citations.

At the level of the single search, we can have little statistical confidence in the individual recall estimates, although our confidence will obviously vary with the size of the recall base for any one search. We can have more confidence in the recall estimate 15/17 than we can in the recall estimate 0/1. However, in the following analyses we are not basing any observations on performance figures (either recall ratios or precision ratios) for single searches. All analyses are based on the grouping together and averaging of

performance figures for several individual searches having some characteristic in common (for example, processed by the same MEDLARS center or falling in the same broad subject area). Our confidence in an average performance figure based on a number of individual recall estimates is obviously much greater than the confidence of any individual estimate.

It is also extremely important to recognize that the recall ratio and the precision ratio for any one search are based on two completely separate document sets which may or may not overlap. The recall set is arrived at by methods extraneous to the system (i.e., articles cited by the requester or found by parallel manual search). It is coincidental if this set happens to overlap the random sample selected from the MEDLARS search printout. Because our two performance figures are based on different document sets, at the single search level we sometimes will get seemingly anomalous results. These anomalies take the form of a positive precision figure and a zero recall figure for the same search. Consider search #18 for which we have a precision ratio of 7/11 (63.6%) and a recall estimate of 0/4 (0). Obviously this is a logical absurdity. Recall cannot be zero if the precision ratio is positive; in this case we know that the search retrieved at least seven relevant articles. However, we can reasonably conclude that the recall for this search was very low. We established for this search a recall base of four documents, known to be in the MEDLARS data base, and discovered that the MEDLARS search retrieved none of them. It is unlikely that these were the only relevant articles not retrieved by the search.

Such anomalies will sometimes be found when we look at performance figures for a single search. However, as already stated, we are not looking at or basing any conclusions on the results of a single search, but only on results averaged over significant groups of searches.

Another problem is presented by the 14 searches having negative results. There are four types of searches involved:

1. The situation in which the requester knows of no relevant articles, and expects none, none are found by parallel manual search, and MEDLARS retrieves nothing. Obviously MEDLARS is behaving perfectly in this case; there are no relevant articles and the system correctly delivers a null response. There is only one search of this type among the 302 analyzed, #56, and this has been given the maximum score of 100% for both recall and precision. In this case 0/0 is taken as equalling 100%.

2. The situation in which the requester knows of no relevant articles, none are found by parallel manual search, but MEDLARS retrieves some that are all judged irrelevant. This is a somewhat worse performance than the first case, but not a complete failure. Take, for example, search #192. To the best of our knowledge there are no relevant articles on the precise topic of interest to the requester. MEDLARS has not behaved perfectly in retrieving five citations, all irrelevant. Nevertheless,

the search serves the same purpose as #56 did. Once the requester has examined the five citations and judged them of no value, he is satisfied that no relevant journal literature exists within the time span of the system. In this case it has cost him a little more effort, but the search proves non-existence as conclusively as did #56. For searches such as this (there are five: #169, #192, #302, #463, #489) we have scored the recall ratio as 0/0 = 100%, but the precision ratio as 0/X = 0.

3. The situation in which we have a valid recall base of relevant articles but MEDLARS retrieves nothing and so informs the requester. Take, for example, search #115. This is MEDLARS functioning at its very worst; the requester knows no relevant literature and the system virtually tells him that no relevant literature exists. However, relevant articles do exist because four were found by parallel manual search. In such cases (#115, #213, #303, #507, #559) we have no qualms in scoring precision as 0/0 = 0, while recall is scored as 0/X = 0.

4. The situation in which the requester knew of no relevant articles and none were found by parallel search. However, MEDLARS confounds us by turning up some relevant items. In such instances we are forced to admit to not having a recall base for the search. Because we have no recall base (although relevant articles exist), we cannot score the search, at least as far as recall goes. There are three searches of this type (#151, #270, and #451) and they have merely been omitted from all statistical tabulations, reducing to 299 the number of searches for which numerical results are tabulated.

Recall ratios and precision ratios cope somewhat less satisfactorily with negative search results than they do with positive. Nevertheless, we have scored the problem searches in a way that appears consistent with the scoring of the positive results.

1.	Recall 0/0 = 1	Precision 0/0 = 1, where no relevant literature exists and MEDLARS retrieves nothing
2.	Recall 0/0 = 1	Precision 0/X = 0, where no relevant literature exists, but MEDLARS retrieves something
3.	Recall 0/0 = 0	Precision 0/0 = 0, where relevant literature exists but retrieval is zero

Average MEDLARS performance for the test requests

When we take the individual performance ratios for the 299 test searches,

and average them, we arrive at the results displayed in Table 4. Over a substantial representative sample of MEDLARS requests, the system was found to be operating, on the average, at 57.7% recall and 50.4% precision. That is, on the average, over the 299 test searches, MEDLARS retrieved a little less than 60% of the total of relevant literature within its base. At the same time, on the average, approximately 50% of the articles retrieved were of some value to requesters in relation to the information needs prompting their requests to the system. Approximately 25% of the articles retrieved were judged of major value by the requesters.

In a sense, the recall ratio based solely on major value articles is a better indicator of system performance, coupled with the overall precision ratio, than the overall recall ratio. On the whole, it is probable that an article judged "of major value" by a requester is one that he would not want to miss, whereas a minor value article is one that he is quite happy to see retrieved but does not really care too strongly about. Viewed in this light, we can say that MEDLARS is retrieving about 65% of the major value articles, accompanied by around 50% irrelevancy.

Table 5 presents a breakdown of the components of the recall base. There were 200 searches for which we were able to get requesters to name relevant articles prior to the conduct of the MEDLARS search. Considering only the recall estimates based on these requester-supplied recall sets, and averaging the results over the 200 searches, we come up with a 58.5% recall ratio. There were 249 MEDLARS searches for which a parallel manual search (almost all were conducted at NLM) uncovered some articles judged relevant by the requester. Considering only the recall estimates based on the recall sets established by parallel manual search, and averaging the results over the 249 searches, we arrive at a 59% recall ratio. Note that these two recall bases are not completely mutually exclusive. In a few searches, an article cited in advance by the requester was also found independently by manual search, and thus appears in both bases (although, of course, it was counted only once when both bases were amalgamated to produce the overall recall ratio).

These two results, achieved on two independently derived recall bases, validate one another and support the validity of the methodology adopted to establish a recall ratio. Virtually, we have two completely separate document sets (one set representing the articles known to requesters at the time they made their requests, and the other established by parallel manual search at NLM) for which, over 299 searches, MEDLARS achieves a recall performance identical within one half of one percent.*

*It is also noteworthy that a virtually identical recall figure was obtained on two recall bases established at different times: (a) the time the request was made, and (b) the time the search results were delivered to the requester.

Moreover, when we consider the "best set" results (derived on the basis of articles found by parallel manual search and articles supplied by the requester but not found by him in Index Medicus), which are the results having the least likelihood of bias, the performance difference is unexpectedly insignificant at 58.1%. In other words, inclusion in the recall base of the articles found by requesters in Index Medicus only results in a 0.4%. improvement in the overall recall ratio for the 299 searches. Since there is no significant difference in the overall recall figures, however we derive them, all further analyses in this report are based on recall ratios for the complete recall base (i.e., all articles cited by requesters and all additional articles found by parallel manual search and later judged relevant by the requesters).

Table 4

Summary of average recall and precision ratios for 299 searches (i.e., omitting three that have no recall base)

Overall precision ratio*	50.4%
Precision ratio based on major value articles only	25.7%
Overall recall ratio (complete recall base)	57.7%
Recall ratio based on major value articles only (274 searches)	65.2%

Table 5

Recall ratio breakdown

Ratio based only on articles cited by requester (200 searches):	58.5%
Ratio based only on articles found by parallel manual search (249 searches):	59%
"Best set" recall ratio, based only on articles cited by requester, but not found by him in Index Medicus, plus articles found by parallel manual search (299 searches)	58.1 %
"Best set" recall ratio, based on major value articles only (271 searches)	65.7%

* Unless otherwise stated, all figures in the tables are calculated by averaging the individual ratios.

Table 6
Reasons for 797 recall failures.

(302 searches were examined, and in 238 of these recall failures are known to have occurred).

Source of Failure	Number of Missed Articles Involved	Percentage of Total Recall Failures Involved	Number of Searches Involved	Percentage of the 238 Searches Involved
Index Language				
Lack of appropriate specific terms	81	10.2%	29	12.2%
Searching				
Searcher did not cover all reasonable approaches to retrieval	171	21.5%	80	33.6%
Search formulation too exhaustive	67	8.4%	31	13.0%
Search formulation too specific	20	2.5%	9	3.8%
"Selective printout"	13	1.6%	7	2.9%
Use of "weighted" terms	2	0.2%	1	0.4%
Other searching failures due to sorting, screening, clerical error	6	0.8%	5	2.1%
TOTAL FAILURES ATTRIBUTED TO SEARCHING	279	35.0%	133	55.9%
Indexing				
Insufficiently specific	46	5.8%	31	13.0%
Insufficiently exhaustive	162	20.3%	100	42.0%
Exhaustive indexing (searches involving negations)	5	0.6%	4	1.7%
Indexer omitted important concept	78	9.8%	61	25.6%
Indexer used inappropriate term	7	0.9%	7	2.9%
TOTAL FAILURES ATTRIBUTED TO INDEXING	298	37.4%	203	85.3%
Computer Processing	11	1.4%	7	2.9%
Inadequate User-System Interaction	199	25.0%	70	29.4%
	868*			

*868 factors contributing to 797 recall failures.

Table 7

Reasons for 3038 precision failures.

(302 searches were examined, and in 278 of these precision failures are known to have occurred).

Source of Failure	Number of Unwanted Articles Involved	Percentage of Total Precision Failures	Number of Searches Involved	Percentage of the 278 Searches Involved
Index Language				
Lack of appropriate specific terms	534	17.6%	58	20.9%
False coordinations	344	11.3%	108	38.8%
Incorrect term relationships	207	6.8%	84	30.2%
Defect in hierarchical structure	9	0.3%	5	1.8%
TOTAL FAILURES ATTRIBUTED TO INDEX LANGUAGE	1094	36.0%	255	91.7%
Searching				
Search formulation not specific	462	15.2%	87	31.3%
Search formulation not exhaustive	356	11.7%	62	22.3%
Searcher used inappropriate terms or term combinations	132	4.3%	31	11.2%
Defect in search logic	33	1.1%	6	2.2%
TOTAL FAILURES ATTRIBUTED TO SEARCHING	983	32.4%	186	67.0%
Indexing				
Exhaustive indexing	350	11.5%	137	49.3%
Insufficiently exhaustive (searches involving negations)	5	0.2%	2	0.7%
Indexer omitted important concept (search involving negations)	1	0.03%	1	0.4%
Insufficiently specific	1	0.03%	1	0.4%
Indexer used inappropriate term	36	1.2%	26	9.4%
TOTAL FAILURES ATTRIBUTED TO INDEXING	393	12.9%	167	60.1%
Inadequate User-System Interaction				
Explicable	464	15.3%	85	30.5%
Inexplicable	39	1.3%	26	9.4%
TOTAL FAILURES ATTRIBUTED TO INADEQUATE INTERACTION	503	16.6%	111	39.9%
Computer Processing	3	0.1%	3	1.1%
Value Judgement	71	2.3%	40	14.4%
"Inevitable" retrieval	4	0.1%	4	1.4%
	3051*			

* 3051 factors contributing to 3038 precision failures.

ANALYSIS OF CAUSES OF RECALL AND PRECISION FAILURES

There are 238 searches in which recall failures are known to have occurred. We know 797 cases of recall failures (i.e., 797 articles that should have been retrieved, because they were judged "of value", but were not) over these 238 searches, and the reasons for all of these failures were analyzed in detail according to procedures mentioned earlier. The results are presented in Table 6.

Likewise, there are 278 searches in which precision failures are known to have occurred, and we know of 3038 cases of precision failures (i.e., 3038 articles that were retrieved although they should not have been because they were judged of no value). All of these precision failures were also analyzed, the results appearing in Table 7.

In studying these tables, one fact must be kept clearly in mind; the figures quoted are not absolute figures for the total number of failures occurring. For example, in Table 7, the figure of 534 precision failures due to "lack of appropriate specific terms" is quoted. This does not mean that in 278 searches only 534 unwanted articles were retrieved because of lack of specificity in the vocabulary. We know of 534 and these exemplify a much larger number of failures of this type.* The meaningful figures are the percentages in this case; lack of specificity in the vocabulary contributed to 17.6% of all the precision failures, and affected the precision performance of 20.9% of all the 278 searches in which precision failures occurred.

Analyses of failures: explanatory notes

We will now consider in detail the various factors contributing to recall and precision failures under each of the system components responsible: indexing, searching, index language, the user/system interface, and computer processing. In these analyses, appear three terms that are given special meaning, and therefore require precise definition, namely exhaustivity, specificity, and entry vocabulary.

Exhaustivity and specificity of indexing

Exhaustivity and specificity are terms that apply both to the indexing of a document and to the preparation of a search formulation for a request. By exhaustivity of indexing we mean the extent to which the potentially indexable items of subject matter contained in a document are in fact recognized in the "conceptual analysis" stage of indexing and translated into the language of the system. For example, consider an article that discusses the use of radioisotope brain scanning in the localization of five different types of lesions. If we completely omit to use an index term that would encompass one of these lesions, we are not indexing the subject matter of this document exhaustively. On the other hand, consider an article that presents case histories on, say, 20 patients. If the indexer includes terms to cover all diseases and

* In actual fact, over 4,000 precision failures in the 278 searches.

abnormalities mentioned, all diagnostic techniques used, and all therapeutic procedures, including specific drugs employed, the subject matter of the document has been covered highly exhaustively.

A high level of exhaustivity of indexing will tend to result in a high recall performance for a retrieval system, but also in a low precision performance. Conversely, a low level of exhaustivity of indexing (i.e., inclusion of "most important" concepts only) will tend to produce a high precision, low recall performance. Exhaustivity of indexing is largely controlled by a policy decision of system management. At NLM, guidelines are given to indexers on the number of subject headings that may be applied to an article. This establishes a general exhaustivity level. Within this established level, the indexers choose terms to express what they consider to be the most important concepts discussed in an article. In these analyses, failure to retrieve a relevant document due to the fact that a particular concept was not indexed is called a recall failure due to lack of exhaustivity of indexing. Similarly, the retrieval of an unwanted document because of inclusion of minor importance concepts in indexing is called a precision failure due to exhaustivity of indexing. It is, then, obvious that there is no "correct level" of exhaustivity in any absolute sense. There should, however, be an optimum level in relation to the types of request made of a particular retrieval system.

Specificity of indexing refers to the generic level at which a particular item of subject matter is recognized in indexing. For example, consider indexing the topic "tetrodotoxin". This could be expressed specifically by a single term TETRODOTOXIN or we could deliberately choose to express this subject precisely by the joint use of two terms TOXINS and PUFFER FISH (recording this decision in our entry vocabulary, as: Tetrodotoxin index under TOXINS and PUFFER FISH). Alternatively, we could index this topic at a higher generic level (i.e. at the level of "toxins produced by fish") by the joint use of a term FISH and a term TOXINS. Climbing one level higher in the generic tree, we could index the topic under a term ANIMAL TOXINS, or ZOOTOXINS, or by the joint use of the term TOXINS and the term ANIMALS. We could, of course, choose to be even more general and index this specific toxin under the very broad term TOXINS.

Obviously, a high level of specificity in indexing will tend to produce a high precision capability in a retrieval system, whereas a low level of specificity will result in a low precision capability. This can be demonstrated by Figure 6. Unless we uniquely define that class of documents dealing with tetrodotoxin, we will never be able to retrieve documents on this subject except as part of a larger class of documents -- in this case, as part of the class "fish toxins", the class "animal toxins", or the class "toxins". The greater the specificity of the indexing (i.e., the smaller the size of the document classes uniquely defined), the greater will be our precision capabilities.

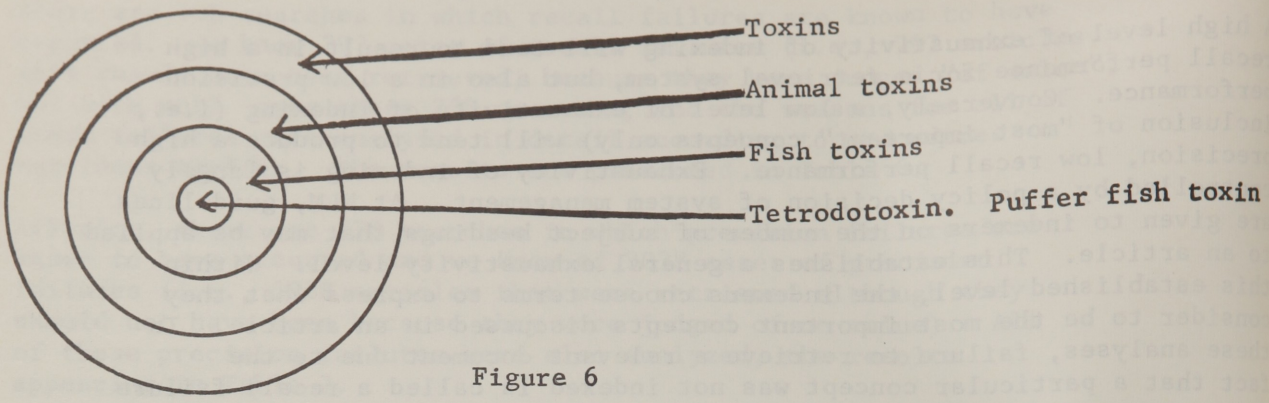


Figure 6

On the other hand, a high level of specificity in indexing will also tend to reduce recall. Reconsidering Figure 6, if we search only the class "tetrodotoxin" for a request on this subject, while those documents retrieved will tend to be highly relevant, we may well be missing some potentially useful items that have been assigned to the more general class "fish toxins" (for example, articles on tetrodotoxin not recognized by the indexer as being on this precise subject, or articles on fish toxins in general which contain substantial information relevant to the subject of puffer fish toxin).

However, to improve our recall, we can compensate for a high level of specificity in indexing by means of our searching strategies. That is, we can broaden the class of acceptable documents by searching at a higher generic level -- in this case by accepting the entire class "fish toxins". However, no variations in searching strategy will ever be able to compensate for lack of specificity in indexing. If we subsume the class "tetrodotoxin" under the general class "animal toxins", there is no searching strategy that will allow us a very high precision search on the subject of tetrodotoxin, since this class of documents is no longer uniquely defined.

Unlike exhaustivity (which is controlled in general by a policy decision of system management, and in particular by decisions made by individual indexers), specificity is governed by the index language -- in the case of NLM, by the characteristics of Medical Subject Headings, and by decisions recorded in such supplementary tools as the MEDLARS Indexing Manual and the Authority File.

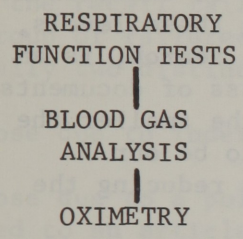
Although the degree of specificity in indexing is governed by properties of the index language, the indexer can in fact index a particular topic at a higher level of specificity than that allowed by the index language. This can either be a deliberate decision made by the indexer (e.g., he chooses to index an article on a number of pulmonary conditions under the term LUNG DISEASES rather than under terms for the specific diseases concerned) or it can be an indexing error (i.e., the indexer chooses a more general term because he is unaware of the existence of the specific term).

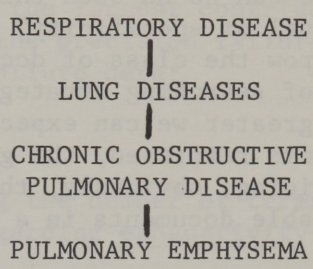
In these analyses, failures due to lack of specificity in indexing (i.e. due to an indexer using a term of a higher generic level than that allowed by the index language) are distinguished from failures due to inherent lack of specificity in the index language.

Exhaustivity and specificity in searching

At the searching stage, the notions of exhaustivity and specificity are much less precise than they are in indexing; in fact, they tend to merge into one another. To take a very simple example, imagine a request for literature on oximetry applied to patients with pulmonary emphysema. This request involves only two facets or categories: the measurement technique facet and the disease facet. If we recognize both facets or categories, and demand their co-occurrence in our search formulation, we are being exhaustive in our formulation.

However, we can recognize each of these facets at any of several levels, as illustrated below:





If we specify that both facets occur at exactly the level of specificity demanded in the request, i.e. we use terms defining precisely the classes "oximetry" and "pulmonary emphysema", we are being fully exhaustive and fully specific in our search formulation for this request. We can, therefore, expect that the group of retrieved documents will, in the main, be highly relevant to the request (i.e. we will achieve a high precision search).

To improve our recall for this request, by pulling in a larger class of potentially relevant documents, we can move in one of two directions. Either we can reduce the exhaustivity of the formulation or we can reduce its specificity. We reduce specificity by moving up in one of the hierarchies, without omitting it entirely. For example, we could move to the more generic class "blood gas analysis" and demand that this co-occur with a term indicating "pulmonary emphysema". Or we could reduce specificity in the disease category and ask, as an example, for the co-occurrence of the class "lung diseases" and the class "oximetry". Of course, we can reduce specificity in more than one category simultaneously. For instance, we could demand co-occurrence of the class "blood gas analysis" and the class "lung diseases". Alternatively, we can broaden our search, with the

object of improving recall, by reducing exhaustivity in the formulation (i.e. by omitting a category entirely). If we asked only for the class "oximetry" we would be searching at a low exhaustivity level for the stated request.

Both high level of exhaustivity and high level of specificity in searching, since they reduce the class of acceptable documents, make for high precision and low recall. Broadening the class of acceptable documents, by reducing specificity and/or exhaustivity in the formulation, will tend to improve recall and reduce precision.

In the analysis, failures to retrieve wanted documents because of a stringent search formulation are characterized as failures due to "formulation too exhaustive" or "formulation too specific". Conversely, failures to hold back unwanted documents, due to a relaxed search requirement, are characterized as failures due to "formulation not specific" or "formulation not exhaustive".

Obviously, there can be no such thing as a "correct level of exhaustivity" or a "correct level of specificity" in searching. Varying these levels, to widen or narrow the class of documents accepted in the search, is an essential part of searching strategy. The larger the class of documents retrieved, the greater we can expect to be our recall; the smaller the class of documents retrieved, the greater we can expect to be our precision (providing, obviously, that we are enlarging or reducing the class of acceptable documents in a sensible fashion).

Entry vocabulary

The importance of an adequate entry vocabulary can be demonstrated by a reconsideration of Figure 6. We said earlier that the class of documents dealing with "tetrodotoxin" can be uniquely defined in the language of the system, or that the class can be subsumed under some larger class, thereby losing its separate identity. We also said that the precision performance of a system is directly controlled by the size of the document classes uniquely defined by the index language (i.e., by the level of specificity in indexing). However, from the point of view of recall, it matters little whether we uniquely define a class or subsume it under some larger class, as long as we record the decision taken. Thus, to return to our toxin example, we can uniquely define the class "tetrodotoxin", we can subsume it under "fish toxins", we can subsume it under "animal toxins", or we can subsume it under "toxins". Whichever decision we make, the class of documents relating to "tetrodotoxin" can be retrieved in subsequent searches, providing we have an entry in our entry vocabulary to tell us where this particular document class has been put, as:

Tetrodotoxin	index under	FISH TOXINS
	or	
Tetrodotoxin	index under	ANIMAL TOXINS
	or	
Tetrodotoxin	index under	TOXINS

Obviously, the precision will deteriorate with the size of the document class retrieved. However, our entry vocabulary leads us to the group of wanted documents, and allows us to retrieve it, however it has been subsumed.

An adequate entry vocabulary is essential to ensure that indexers and searchers consistently use the same terms, or term combinations, to describe identical items of subject matter. A rich entry vocabulary will include all words and phrases used in documents to describe notions that have been recognized in the conceptual analysis stage of indexing and translated into the language of the system. It will also include words and phrases used in requests to describe notions about which literature exists in the system. The quality of an entry vocabulary can substantially affect the recall performance of an information retrieval system.

Recall and precision failures attributable to the indexing subsystem

From Tables 6 and 7, it can be seen that the indexing subsystem contributed to 37% of the recall failures, and was in fact the largest contributor to this group of failures, but to only 13% of the precision failures. There are really two distinct types of indexing failure here:

1. Those due to indexer errors.

2. Those due to a policy decision governing the number of terms assigned to an article (i.e., the policy regarding exhaustivity of indexing).

Indexer errors are themselves of two types: (a) omission of a term or terms necessary to describe an important topic discussed in an article, and (b) use of a term that appears inappropriate to the subject matter of the article. Omissions will normally lead to recall failures, while use of an inappropriate term can cause either a precision failure (the searcher uses this term in a strategy and retrieves an irrelevant item) or a recall failure (the searcher uses the correct terms and a wanted document is missed because labeled with an incorrect term). Use of inappropriate terms (i.e., sheer misindexing) is negligible in MEDLARS, contributing to about 1% of the precision failures and 1% of the recall failures. The misindexings that do occur appear not to be errors of carelessness. Rather they appear due to the general misuse of a particular term at some point in time. For example, RADIOISOTOPE SCANNING was used indiscriminately for any radioisotope monitoring operation, whether or not scanning was involved.

Indexer omissions, on the other hand, contribute, significantly, to almost 10% of all the recall failures, and were wholly or partly responsible for at least one missed article in 25% of all the searches in which recall failures occurred. These omissions are fairly gross errors and cannot be attributed primarily to a policy decision governing indexing

exhaustivity.* We have distinguished recall failures due to indexer omissions from recall failures due to lack of exhaustivity of indexing, as follows:

1. Indexer omission: a topic that appears central to the subject under discussion in the article is not covered at all in the indexing. It is felt that the omitted topic is so important that it should be covered even in "non-depth" indexing.

2. Lack of exhaustivity: an item of subject matter treated peripherally in the article is not covered in the indexing. The topic is not crucial to the article, and was presumably excluded in favor of other topics due to general policy regarding the average number of terms to be assigned, and perhaps also to the short time period allowed for indexing (an experienced indexer at NLM will index 40-50 articles per day).

The following are two examples of indexer omissions discovered in test searches:

20 An article on the effect of visual deprivation on growth of the visual cortex in mice was indexed under CEREBRAL CORTEX and VISION and DARKNESS, but should also have been indexed under GROWTH and SENSORY DEPRIVATION. It is from a depth-indexed journal and was not retrieved in a search relating to growth in the nervous system, although regarded as of major value.

39 A major value article on reversible sterility following discontinuance of medroxyprogesterone was not retrieved because the crucial term STERILITY, FEMALE was not applied. The search was on prolonged amenorrhea and infertility following discontinuance of oral contraceptives.

Unfortunately, if an important term is omitted from the indexing of an article, this item is likely to remain unretrieved in a number of searches to which it is highly relevant. For example, test search #1 and test search #527 both relate to the crystalline lens. An important Science article on induction, particularly the induction of the crystalline lens in salamanders, was not retrieved in either search, although of major value, because the term LENS, CRYSTALLINE was not applied. Moreover, this type of error is likely to remain undetected in the normal operations of the system.

*To test this, ten articles not retrieved because of indexer omissions were submitted to the Index Section for re-indexing. In all cases, the formerly missed term was applied in the second indexing.

Although a certain number of indexing omissions are to be expected under the pressures of a tight production schedule, some of those occurring are difficult to excuse, particularly when a term appearing in the title of an article is omitted. For example, search # 45 relates to electrical brain stimulation in elicitation of species-specific behavior. One of the missed articles, from a 1966 issue of Behavior, was not indexed under ELECTRIC STIMULATION even though entitled "Behavioral effects of electrical stimulation in the forebrain of the pigeon". Likewise, search # 99, on phosphorus or phophates in the brain, missed an article entitled "Incorporation of ortho [32p] phosphate into the subcellular fractions of developing rat brain" because it had not been indexed under BRAIN or any other term indicating cerebral involvement.

A significant number of these cases of indexer omission can be attributed to the fact that no Mesh term exists for the missed notion, and there is nothing in the entry vocabulary to say how the topic is to be indexed. As a result, the indexer either omits the topic entirely or indexes it much too generally. This type of failure was found, for example, in search # 190, which relates to deiodination of thyroxine. A major value article, unretrieved, deals with flavin photodeiodination of thyroxine. There is no Mesh term for "photodeiodination", or indeed for "deiodination", and there is nothing in the entry vocabulary to say how this concept is to be indexed. Consequently, the notion was completely ignored in indexing, although it might reasonably have been translated into IODINE. Similar, but much more drastic, failures occurred in search #102 and search # 177. Search # 102 sought articles on hemodynamic analysis using Fourier series. Many major value articles (for example, on the Fourier analysis of vascular impedance) were missed. Despite the fact that "Fourier analysis" or "Fourier series" appeared prominently in these articles, this aspect was ignored by the indexers. Presumably these failures are due partly to lack of an adequate entry vocabulary. There is no specific term for Fourier series. Despite this, relevant articles could have been retrieved were there an entry in the entry vocabulary to tell indexers and searchers that "Fourier analysis" is to be subsumed under the broader term MATHEMATICS or whatever other term is chosen to express the topic. The same kind of failure occurred in a very unsuccessful search (# 177) on premature rupture of the fetal membranes. Because no specific term exists, this topic was not covered in a number of articles in which it is discussed centrally.

On the surface, it may appear strange that there should be so few instances of misindexing (that is, use of inappropriate terms) but so many cases of indexer omissions. The explanation may be simple. The work of the inexperienced indexers is scanned ("revised") by senior indexers. Usually the inappropriate terms will stand out quite clearly and are easily corrected in this revision process. However, omissions are not so readily detected by the reviser because this would involve a careful examination of the article. Consequently errors of omission are more likely to creep through than examples of sheer misindexing.

Failures due to exhaustive indexing or to lack of exhaustivity

As previously mentioned, the more exhaustively we describe (by means of index terms) the subject matter of documents, the greater will be the recall potential of the system. Conversely, because of the inverse relationship between recall and precision, the more exhaustive the indexing, the more precision failures are likely to occur. This is partly attributable to the greater potential for false term coordinations, and partly to the fact that exhaustive indexing will cause retrieval of articles in response to requests to which they relate very weakly. In the operation of any retrieval system we are likely to find recall failures caused by indexing that is not sufficiently exhaustive. At the same time, we will discover precision failures due primarily to the fact that exhaustive indexing has brought out articles on topics for which they contain very little information.

This is exactly what happens in MEDLARS. Twenty percent of the recall failures are attributed to lack of exhaustivity of indexing, while 11.5% of the precision failures are caused largely by exhaustive indexing. It may be worthwhile at this point to recapitulate on the MEDLARS situation in relation to indexing exhaustivity. At the present time, there are actually three levels of exhaustivity within MEDLARS. Since September 1964, the complete list of journals indexed has been divided into two parts: "depth" and "non-depth". Articles from "depth" journals (about one third of all the 2400 journals regularly indexed) are presently indexed at an average of about ten index terms per article, while the non-depth articles are indexed at an average of slightly less than four terms per article. The overall average for depth and non-depth is about 6.7 terms per item. In addition, some of the terms assigned to both depth and non-depth articles are chosen to be the headings under which entries for the articles will appear in Index Medicus. Only the terms representing the most important topics discussed in an article are chosen as print or IM terms. Thus, the print terms can function as weighted index terms. At present, there are approximately 2.6 print terms per article

A small experiment was conducted to determine how the proportions of depth to non-depth articles had varied over the years, and how the average number of index terms assigned had varied. One hundred citations were selected at random from the author index to the 1964 Cumulated Index Medicus, and equal samples were also drawn from the 1965 and the 1966 author indexes. Citation printouts for these items were obtained, and the number of terms assigned to each was counted. The results are presented in Table 8. Approximately 42% of the 145,000 1964 citations were input from journals now on the "depth" list, and 58% from journals on the "non-depth" list. It can be seen that, although there was no formal distinction between "depth" and "non-depth" at the time, the indexers appeared to be aware of some overall quality distinction between the two, assigning an average of 7 terms to articles from the present "depth" journals and an average of 5.9 terms to articles from the present "non-depth" journals.

For the 171,000 1965 citations, the proportions changed: now around 54% of the total of input citations were from depth journals.

Moreover, the formal distinction between "depth" and "non-depth" journals led to a widening in the term-assignment gap, articles from the former being assigned an average of 7.6 terms as compared with an average of 4.2 terms for articles from the latter.

For the 164,000 1966 citations, the proportion of depth to non-depth showed a further increase to 58:42. Now, however, a general trend towards more exhaustive indexing, both for depth and non-depth journals, meant that the gap between the two had closed slightly. As compared with the 1965 citations, the 1966 "depth" articles showed an average term increase per item of ·7 (7.6 to 8.3), while the "non-depth" showed an average term increase of 1.7 (4.2 to 5.9). Since in 1967 the proportion of depth to non-depth is again increasing, we can estimate from Table 8 that, during the period of the present test (i.e., August 1966 - July 1967), the MEDLARS file was divided in roughly the proportion of 55% from depth journals, 45% from non-depth.

Recall failures due to indexing of insufficient exhaustivity

One example is adequate to illustrate a typical recall failure of this type. Search # 535 relates to the transmission of viral hepatitis by parenteral inoculations of materials other than blood or blood products or during venipuncture. One major value article that was not retrieved deals with hepatic inflammation in narcotic addicts. The fact that viral hepatitis is transmitted by contaminated injection equipment was mentioned in the text but was not covered by the indexing.

Recall failures due to lack of exhaustivity have been taken very literally in this evaluation. Apart from the failures due to indexer omissions, we have attributed to this cause any recall failures in which the relevant article deals in some way with the subject of the request, but this aspect was not covered in the indexing. In some cases, the "relevant" section of an article is very minor indeed, and it could only be covered by an extremely high (and probably uneconomical) level of exhaustivity. For example, search # 49 deals with the relationship between hypertension and speech disorders. The only known relevant article, within MEDLARS, is a 25-page review on vocal behavior, which contains only one paragraph on vocal behavior in patients with high blood pressure. Similarly, a search on potassium shifts in isolated cell preparations (# 34) missed a general review article (Physiological Reviews) in which sodium and potassium fluxes are dealt with very briefly.

Lack of exhaustivity of indexing will normally cause recall failures. In searches involving negations, however, it can lead to precision failures. This occurred in a search on life islands in relation to humans (# 123) which was searched on GERM-FREE LIFE with animal terms negated. Unexpectedly this retrieved a review article on germfree

Table 8

Variations in indexing treatment of "depth" and "non-depth" journals based on three random samples of 100 articles selected from 1964, 1965, and 1966 Cumulated Index Medicus.

Year	Proportion of "depth" to "non-depth" articles	Average # of index terms assigned		Range		Mode	
		D	ND	D	ND	D	ND
1964	42:58	7.0	5.9	2-14	2-17	8	3
1965	54:46	7.6	4.2	3-15	2-15	3	4
1966	58:42	8.3	5.9	3-18	2-16	6	4

animals, which had been indexed under GERM-FREE LIFE, REVIEW and MICROBIOLOGY, but not under any animal terms.

Precision failures due to exhaustive indexing

Exhaustive indexing will contribute to a small proportion of the precision failures in certain searches and will be largely responsible for all the precision failures in others. For example, one of the irrelevant items retrieved in a search on the crystalline lens in vertebrates (# 1) deals with the correlation between mast cells and the histamine content of the eye in cattle. It was indexed under LENS, CRYSTALLINE even though only one item of data on the lens in presented. Search #13, on blood or urinary steroids in human breast or prostatic neoplasms, retrieved two types of irrelevant article due to exhaustive indexing:

1. Articles in which the required neoplasm aspect is barely mentioned (for example, an article indexed under BREAST NEOPLASMS deals with plasma androgens in women and discusses a number of patients, only one of whom had breast cancer).

2. Articles in which the urinary aspect is very slight (PROSTATIC NEOPLASMS and URINE retrieved, for example, a single case report on prostatic cancer in which a urinary hormone assay value is presented in a table).

In other searches, exhaustive indexing had quite a drastic effect on precision. A search on the action of chloramphenicol (# 46) retrieved 339 citations. In about 50% of the articles cited, the reference to chloramphenicol is very slender (e.g., it is used as an incubation medium in a bacterial study). The precision ratio for this search was only 20%.

Search # 148 relates to the tubular secretion of creatine. Only 15% of the 500 retrieved citations were judged of any value. Many of the retrieved articles contain very little directly on creatine; for example, they may refer to a creatine value obtained in a routine kidney function test. In this case, the searcher could have raised precision to 25%, without recall loss, by requiring that CREATINE AND CREATININE be a print (Index Medicus) term.

Some of the worst failures of exhaustive indexing appear due to some rather peculiar policies adopted in indexing. It seems that any article that merely mentions the word "computer" ("Calculations were conducted on an IBM 7094 computer") is likely to be indexed under some data processing term. Consequently it becomes very difficult to conduct a successful search on a specific biomedical application of computers. Search # 47 deals with computer recognition of cells, but about half of the 79 retrieved articles simply mention that a computer was used in calculation, and some (e.g., an article on the flight control system of grasshoppers) are very far away from the search topic.

Geographical terms are also used very loosely, and not strictly for material with a definite geographical connotation. This can lead

to disappointing results in a legitimate regional search. A search on medicine in Minnesota (# 55) retrieved 235 articles, but about 50 of these contain no real reference to Minnesota. The term MINNESOTA has, for example, been applied to case studies emanating from the Mayo Clinic.

Effect of exhaustivity levels

Seventy-three searches were selected for additional analysis to determine the effect of indexing exhaustivity on performance. This group of searches included 39 recall failures due to lack of exhaustivity and 69 precision failures due to exhaustive indexing. It was noted that the average number of index terms assigned to the 39 missed articles was 7.6, while the average number assigned to the 69 retrieved but unwanted items was 11.8.

It was also possible to derive figures indicative of performance variations caused by differences between depth and non-depth indexing. The combined random sample forming the overall precision base for the test searches consisted of 6491 articles, 4884 from depth journals and 1607 from non-depth. The overall precision ratios for these components, calculated by the average of numbers, was 2386/4672 (51.1%) for the depth articles and 553/1266 (43.7%) for non-depth.

The recall difference between depth and non-depth indexing was calculated on a sampling basis. Twenty searches having large requester-supplied recall bases were selected. The combined recall base consisted of 225 articles, 201 from depth journals and 24 from non-depth. The overall recall ratio for the depth journals, calculated by the average of numbers was 141/201 (70.1%). The overall recall ratio for non-depth was 13/24 (54.2%).

A further analysis was performed to determine what type of performance could be expected if the much lower level of exhaustivity adopted for Index Medicus (about 2·6 terms per article) was also the only level available in the retrospective search system. Because print terms are asterisked in the computer printout of index records (tracings), it was possible to determine the recall and precision results for a number of the test searches based only on these terms. The analysis simply involved the matching of tracings, for both recall and precision base documents, against the search formulation prepared for the request, and the counting of all articles that would have been retrieved had the indexing consisted only of the print terms. A total of 111 searches were analyzed. However, in the case of 23 of these searches the comparison was unreasonable because the search formulation used as a coordinate a term (a check tag such as HUMAN, or a technic term such as MICROSCOPY, ELECTRON), that would either never be used as a print term or would be used very rarely. Consequently, in nine of these searches no documents would have been retrieved on print terms only, while only a handful would have been retrieved in the other 14 searches. These 23 searches were therefore eliminated, all further analysis being conducted on the remaining 88.

The total combined recall base for these 88 searches comprised 633 articles. In the searches on the full MEDLARS indexing, 382 were retrieved, giving an overall recall ratio (by the average of numbers rather than the average of ratios) of 60%. Only 280 of the articles would have been retrieved on the print terms however (i.e., a recall ratio of 44%). In other words, the much reduced exhaustivity of Index Medicus indexing led to the loss of one relevant document in every four.

The total combined precision base for the 88 searches consisted of 1716 documents, of which 890 were judged relevant (i.e., a precision ratio of 52%). Only 783 of these articles, of which 466 (60%) were relevant, would have been retrieved on the basis of the print terms only. In other words, searching on Index Medicus terms alone would have lost 344 of the 890 relevant documents (49%) with a compensatory gain in precision, because 509 of the 826 irrelevant items (61%) would also have been filtered out.

Summarized, the figures are as follows:

	Recall ratio	Precision ratio
Complete indexing	60%	52%
Index Medicus indexing only	44%	60%

These figures, of course, demonstrate the customary effect of variations in indexing exhaustivity: the more terms used, the greater will tend to be the recall but the lower the precision; the fewer, more selective the terms used, the lower will tend to be the recall and the higher the precision.

We can now look at this combined evidence ralating to recall and precision performance as related to exhaustivity of indexing in MEDLARS. The data are presented in Table 9.

Table 9

Effect of indexing exhaustivity on retrieval performance in MEDLARS

	Recall ratio	Precision ratio
Depth indexing (10 terms approximately)	70%	51%
Non-depth indexing (4 terms approximately)	54%	44%
Index Medicus (2.6 terms approximately)	44%	60%

All other things being equal, we would expect that these results would show a strict inverse relationship between recall and precision. This is apparent in the relationship between the Index Medicus terms and the depth indexing. However, the results for the non-depth indexing do not correspond to the expected pattern: both recall and precision for the non-depth articles are lower than for depth articles. This is due to the fact that the non-depth articles are indexed not only with fewer terms, but also with more general terms. We will return to this matter, and discuss its implications, later.

Two small experiments were conducted to determine (a) whether re-indexing "in depth" of unretrieved non-depth articles would appreciably improve recall, and (b) whether re-indexing of non-retrieved depth articles, again in depth, would allow their retrieval. Eighteen non-depth articles that were unretrieved in a number of test searches (although the failure was not necessarily attributed primarily to non-exhaustive indexing) were re-indexed according to "depth" standards. Sixteen of the articles were indexed by three separate indexer-reviser pairs, and the other two articles were indexed by two separate indexer-reviser pairs. The average number of terms assigned in the original non-depth indexing was 5, while the average term assignment for the re-indexing was 11.2.

Seven of the eighteen articles (38.9%) would have been retrieved on the basis of at least one out of three versions of the re-indexing, while six of the eighteen (33.3%) would have been retrieved on the basis of at least two of the three versions.

The failure to retrieve five of the eighteen articles was originally attributed to non-exhaustive indexing, and all five of these would have been retrieved by at least one of the three versions of the re-indexing, while 4/5 (80%) would have been retrieved on the basis of at least two of the three versions.

Thirteen of the failures were not originally attributed to nonexhaustive indexing, but two of these articles could in any case have been retrieved by the additional index terms assigned in the depth indexing. In other words, exhaustive indexing has a certain "fail-safe" property which, under certain conditions, will compensate for other system failures, such as inadequate searching strategies.

The above experiment suggests that depth indexing of non-depth articles might allow retrieval of between 30% and 40% of the relevant non-depth articles that are presently unretrieved in MEDLARS searches.

The second small experiment involved the re-indexing of thirteen articles from depth journals that were not retrieved because of indexing of insufficient exhaustivity. These articles were selected at random from all the recall failures of this type. It was not to be expected that this reindexing would have much effect on recall. The articles were originally indexed in depth at an average of 7.2 terms per item. The omitted topics are comparatively minor aspects of the articles, and only indexing at a substantially higher level of exhaustivity would be likely to affect recall significantly. The test confirmed this. The re-indexing was done by one experienced indexer at a slightly higher level of 9·1 terms per item, but the topic originally omitted from the indexing was added to only two of the thirteen articles.

Failures due to lack of specificity in indexing

Only one precision failure was attributed to the failure of an indexer to use the most specific MeSH term available. However, 5.8% of all the recall failures were due to lack of specificity in indexing. In MEDLARS, lack of specificity and lack of exhaustivity of indexing are both closely related to policy regarding indexing depth (i.e., the average number of terms assigned). Articles from non-depth journals tend to be indexed in general terms. For example, search # 64, on spina bifida and anencephalus, failed to retrieve a number of non-depth articles because they were indexed more generally under ABNORMALITIES. In depth indexing, the specific malformations would have been indexed. Below are two further examples of searches in which recall failures occurred through non-specific indexing:

4 Tissue culture of lung or bronchial neoplasms. A major value article describes the effect of various steroid hormones on 12 cell lines, including adenocarcinoma of the lung. The cell lines were indexed very generally under NEOPLASMS.

57 E coli and lipopolysaccharides. A detailed major value article on the chemistry of E coli polysaccharides, from a depth journal, was indexed under POLYSACCHARIDES rather than the specific polysaccharides mentioned. A second article, dealing specifically with E coli lipopolysaccharides, was indexed under POLYSACCHARIDES, BACTERIAL. Both articles deserve the term LIPOPOLYSACCHARIDES.

General observations on exhaustivity and specificity of indexing within MEDLARS

1. Indexing exhaustivity is, of course, a relative matter. We have already mentioned the article on mast cells and histamine content of the eye in cattle, which was retrieved on the term LENS, CRYSTALLINE although only one item of data was given on this topic. The requester regarded it as of no value in relation to his research on the crystalline lens in vertebrates. Therefore, we must say that exhaustive indexing was largely responsible for this precision failure and, judged in relation to this request, assignment of the term LENS, CRYSTALLINE was unjustified. Visualize, however, a second, much more specific request for articles presenting data on the histamine content of the crystalline lens. In response to this request, the above article is highly relevant and may in fact be one of the few articles in which measured values are quoted. Judged in relation to this request, the term LENS, CRYSTALLINE is completely justified.

2. On the whole it is better to err on the side of exhaustive indexing. It is difficult to retrieve an article on X if X has not been covered in the indexing of the item. On the other hand, within MEDLARS the searcher has a limited capability for reducing the exhaustivity by searching only on Index Medicus terms, and thus improving precision for any one search (although inevitably losing some recall). Alternatively, the searcher may use what is in effect a weighting device by demanding that the key term of a request (LENS, CRYSTALLINE in the above example) must be a print term, but not putting any such restriction on the coordinate terms. This procedure, which is likely to improve precision with rather less drastic effects on recall, will be discussed later in the section on searching failures.

3. The artificial separation of all MEDLARS journals into depth and non-depth appears, from the detailed search analyses, to lead to indexing anomalies that can cause both recall and precision failures. Although many of the articles from non-depth journals seem somewhat superficial and repetitive, others are very substantial papers which, because of a general policy decision, are indexed completely inadequately. On the other hand, half-column letters in Lancet are sometimes assigned 15-20 terms, and are thus retrieved in searches to which they contribute little or nothing. A policy of treating each article on its own merit, whatever journal it comes from, would reduce such seeming anomalies.

4. The indexing policy with regard to review articles appears to be particularly suspect. Review articles are indexed "non-depth" on the grounds that the material reviewed "was probably indexed in depth in the original". This is hard to justify on a number of grounds:

(a) Some of the "reviewed" literature predates MEDLARS.

(b) A good reviewer may present data in new relationships not revealed by the original articles.

(c) A review article may contain one of the most substantial discussions anywhere in existence on a comparatively rare subject. This occurred, for example, in the search on aspergillosis of the orbit (# 268), a topic upon which there is comparatively little relevant literature. One of the major contributions to this subject is the eight relevant pages contained in a review article on fungal diseases of the eye. Yet this article was not retrieved, and could not be retrieved except by a very broad generalization of the search, because indexed only under EYE DISEASES and MYCOSES.

Similar examples occurred in many other searches. Search # 495 relates to "postural intolerance" in space flight and its parallel in prolonged bed rest. A review article on "medical problems of weightlessness" deals largely with the specific topic of interest. Although of major value, it was not retrieved by MEDLARS because indexed only under WEIGHTLESSNESS, HUMAN and REVIEW. Likewise, a major value review on disseminated interstitial lung diseases was missed in a search on "diffuse lesions of the lung" because indexed generally under LUNG DISEASES/DIAGNOSIS, HUMAN, REVIEW and THORACIC RADIOGRAPHY, although it deals substantially with pulmonary fibrosis, histiocytosis, scleroderma, sarcoidosis, and pneumoconiosis.

5. From the point of view of machine retrieval, the policy of indexing non-depth articles in general terms is indefensible. To quote but one example, in the analysis of search # 531 an article from a non-depth journal (Poultry Science), entitled "Role of streptococcus faecalis in the antibiotic growth effect in chickens" was examined. Found by manual search, but missed by MEDLARS, it was indexed only under EXPERIMENTAL LAB STUDY, INTESTINAL MICROORGANISMS and POULTRY.

Use of the general term INTESTINAL MICROORGANISMS for the specific organism implicated is inexcusable. On the basis of this indexing, one could not reasonably expect the article to be retrieved in response to a request on "streptococcus faecalis in poultry" or one on "effect of penicillin on streptococcus faecalis" or even one on "antibiotic growth effect in poultry", to all of which specific topics it is highly relevant. In fact, on the basis of the indexing one could only reasonably expect to retrieve it in a search on intestinal microorganisms of poultry, to which general subject it is indeed a slight contribution.

It is always a mistake to index specific topics under general terms. In the above example, use of the term STREPTOCOCCUS FAECALIS would allow retrieval of this item in response to a request involving this precise organism. On the other hand, the article could still be retrieved in a more general search relating to intestinal microorganisms, because the searcher is able to "explode" on all bacteria terms. The article could have been indexed very adequately under five terms: POULTRY, PENICILLIN, STREPTOCOCCUS FAECALIS, GROWTH, and EXPERIMENTAL LAB STUDY. As presently indexed, it is difficult to visualize a single retrospective search in which it would be retrieved and judged of major value. In other words, this citation and others indexed in such general terms are merely occupying space on the citation file.

Evaluation of indexing as part of the input subsystem

It is difficult to evaluate the components of the overall input subsystem in MEDLARS. We must usually accept as the indexer input the index terms now recorded on the search file. However, there is no guarantee that the terms on the citation file are actually the terms assigned by the indexers. Some terms, for example, could be omitted or changed in the computer input (flexowriter) operations; others could be lost through imperfect file maintenance procedures. Unfortunately, for the evaluator, indexer data sheets and flexowriter hard copy are destroyed once the citations have been input to the system. It thus becomes impossible to separate true indexer errors from subsequent errors of computer input and file maintenance. However, to allow a sample check, all indexer data sheets and all flexowriter hard copy for the 164,000 articles input in 1966 were retained by the Evaluator. Four of the recall failures due to indexer omission (this being the type of failure most likely to be misattributed) involving articles in the 1966 Cumulated Index Medicus were selected for the spot check. Searching of the stored data sheets was a laborious process, because they are sorted only into broad groups by month of input. However, eventually both data sheets and flexowriter hard copy for the four test articles were located. Examination of these showed that the indexer had in fact omitted the important term in three of the four cases, but in the fourth case the term was included on the data sheet. Further examination revealed that it was also included on the flexowriter proof copy, and that the term appeared with the citation in the December 1966 issue of Index Medicus and again in the 1966 Cumulated Index Medicus. The fact that a citation printout now reveals that this term (PARATHYROID GLANDS) is no longer carried among the tracings for the article, indicates some subsequent failure of file maintenance procedures.

On the basis of this small test, we are forced to conclude that perhaps 25% of the failures attributed to indexer omissions in fact occurred later than the indexing stage. In other words, the true proportion of recall failures due to indexer omissions may actually be about 7% rather than the 9.8% shown in Table 7. Failures due to lack of exhaustivity, on the other hand, are almost certainly due to deliberate decisions made by

the indexer within the constraints of indexing policy. It is unlikely that a failure attributed to lack of exhaustivity is due instead to loss of a term in flexowriter input or loss of a term through file maintenance. We are therefore confident in the figure of 20% recall failures due to nonexhaustive indexing.

Recall and precision failures attributable to the searching subsystem

Considering recall and precision failures together, the searching subsystem is the greatest contributor to all the MEDLARS failures, being at least partly responsible for 35% of the recall failures and 32% of the precision failures. We can distinguish three types of searching failure:

1. Pure errors involving the use of inappropriate terms or the use of defective search logic.

2. Failures due to the levels of specificity and/or exhaustivity adopted in searching strategies.

3. Recall failures due to the fact that the searcher did not cover all reasonable approaches to the retrieval of relevant articles.

Recall losses resulting from failure to cover all reasonable approaches to retrieval

In the recall analysis, 21.5% of all the failures were attributed to the fact that the searcher did not cover all reasonable approaches to the retrieval of literature relevant to the request. In other words, 21.5% of the missed relevant articles could have been retrieved on terms or term combinations which, the author feels, the searcher might reasonably have been expected to use in the search formulation. This "failure to cover all reasonable approaches" in searching was a major contributor to recall losses in the 302 searches analyzed, being second only to failures of user-system interaction, which were responsible for 25% of the recall losses.

There are really two categories of failures of this type:

1. Failure to use one particular relevant term, or term combination, in a formulation which otherwise reflects the complete interests stated in a request.

2. Failure to cover a complete aspect of the request as stated by the requester.

The first type has less drastic results than the second, but nevertheless can substantially reduce the recall ratio for a search, as the following

illustrate:

19 In a search on nervous tissue culture as affected by electrical stimulation, and certain other variables, the searcher did not explode on NERVOUS SYSTEM in coordination with TISSUE CULTURE and terms for the specific factors of interest. A directly related article could have been retrieved on CEREBRAL CORTEX and TISSUE CULTURE and ELECTRIC STIMULATION.

34 In a search on potassium shifts in isolated cell preparations, no use was made of the term CELL MEMBRANE PERMEABILITY. Used in conjunction with POTASSIUM or POTASSIUM CHLORIDE, it would have brought out several major value articles.

79 In a search on oral manifestations of neutropenia, the only terms used to express "oral manifestations" were ORAL MANIFESTATIONS, DIAGNOSIS and various anatomical terms. No attempt was made to search on term combinations describing particular possible manifestations (e.g., AGRANULOCYTOSIS and STOMATITIS).

More drastic failures occur when the searcher omits a complete aspect of a topic that is explicitly stated in the request. This type of failure is particularly prone to occur with fairly long, multifaceted request statements. Whether the searcher overlooks the aspect through careless reading, or deliberately ignores it, it is difficult to establish. An example of this type of failure occurred in search # 174, which relates to testicular biopsy in cases of infertility. One aspect of interest (the effect of surgical and hormonal therapy on sperm count, testicular morphology, and fertility) was completely omitted from the formulation, contributing to the low recall ratio of 3/11 (27.3%).

Search # 188 responds to a request in two parts: (1) filaria parasites of primates, and (2) insect vectors of filaria, life cycles and transmission of the filaria. Only the first aspect was covered in the formulation. The second, which is not restricted to primates, was ignored, leading to a recall ratio of only 3/8 (37.5%).

Precision failures due to searching on inappropriate terms or term combinations

Whereas omission of appropriate terms, from a search formulation, will lead to recall failures, use of inappropriate terms or term combinations will cause precision failures. The author attributes 4.3% of the precision failures to this cause. A few examples are given below:

47 Computer recognition of cells. One strategy involved the coordination of CYBERNETICS with all cell terms. CYBERNETICS is inappropriate to a request on cell recognition, which is essentially a pattern recognition problem. It caused retrieval of articles on cells as cybernetic systems, and was responsible for about one third of the irrelevancy in this search.

133 Functions of medical schools, doctors and health agencies in family planning. This search retrieved 285 citations, of which about 240 were completely irrelevant. The searcher used some quite extraordinary term combinations, including OBSTETRICS and PREGNANCY and PHYSICIAN-PATIENT RELATIONS and PREGNANCY, which retrieved many articles on maternal care.

177 Use of the term ABRUPTIO PLACENTAE (separation of the placenta) is inappropriate to a search on premature rupture of the fetal membranes. It was responsible for 80% of the total retrieval (50), and all but one of these items were irrelevant.

Inappropriate term combinations tend to occur with fairly complex search formulations in which a list of terms in a logical sum (or) relation is anded with a second list of summed terms. While the overall strategy may appear sensible, some of the combinations resulting are irrational in relation to the request. This appears to be the cause of the OBSTETRICS and PREGNANCY type of combination encountered, for example, in search # 133.

Recall and precision failures due to variations in exhaustivity of the formulation.

As previously mentioned, varying the exhaustivity and/or specificity of the formulation is an essential part of searching strategy. In fact, the central problem of searching is the decision as to the most appropriate level of specificity and exhaustivity to adopt for a particular request. The less specific and exhaustive the formulation, the more documents will be retrieved, recall will tend to increase and precision to decrease. The more specific and exhaustive the formulation, the fewer documents will be retrieved, recall will tend to deteriorate and precision to improve. For each particular request, we must decide in which direction to go. In other words, how near to 100% recall does the requester really want to approach, bearing in mind that the closer we get to this figure the more documents we are likely to retrieve and the lower is likely to be the precision of the search.

An exhaustive search formulation is one that demands the co-occurrence of all the notions asked for, in some relationship, by the requester (although not necessarily at the level of specificity stated in the request). Consider search # 115, which concerns various specific intestinal microorganisms causing diarrhea or dysentery in cases of protein deficiency or Kwashiorkor This request involves a relationship between three separate notions:

1. Certain specific intestinal microorganisms.

2. The disorder of diarrhea or dysentery.

3. The disorder of protein deficiency or Kwashiorkor.

The searcher was fully exhaustive in the formulation, allowing an article to be retrieved only if:

1. it had been indexed under the term PROTEIN DEFICIENCY or the term KWASHIORKOR, and

2. it had been indexed under a term indicating the involvement of some microorganism , and

3. it had been indexed under a term indicating diarrhea or dysentery.

With an exhaustive formulation such as this, we can expect high precision. That is, most of the articles retrieved are likely to be relevant. On the other hand, our strategy may be too exhaustive; it may be asking too much to expect a relevant article to have been indexed under all of the notions demanded by the requester. This was exactly the case in search # 115, which retrieved nothing, although relevant literature exists and some could have been retrieved with the less exhaustive strategy:

PROTEIN DEFICIENCY

or and diarrhea terms

KWASHIORKOR

Exhaustivity of the search formulation is obviously related to the coordination level (i.e., the number of index terms required to co-occur before an article can be retrieved), but there is no strict one-to-one relationship between exhaustivity and coordination level. For example, PROTEIN DEFICIENCY and DYSENTERY and INTESTINAL MICRO-ORGANISMS is a three-term coordination that is exhaustive in that it covers all the related notions demanded by the requester, but so also does PROTEIN DEFICIENCY and DYSENTERY, BACILLARY, which is a two-term coordination. Moreover, by varying the coordination level, we may be varying the specificity rather than the exhaustivity of the search.

For example, consider a request for "metastatic fat necrosis as a complication of pancreatitis". The formulation PANCREATITIS and NECROSIS is exhaustive in that it asks for the co-occurrence of the two notions specified. The three-term coordination PANCREATITIS and NECROSIS and ADIPOSE TISSUE is merely more specific in relation to the request.

Exhaustive search formulations were responsible for 8.4% of the recall failures and nonexhaustive search formulations were responsible for 11.7% of the precision failures. Some further examples follow:

Exhaustive formulations

217 The request is for "influence of the styloid process on facial and head pains". The searcher required that some term indicating "face" or "head" be present, as well as a term indicating "pain" and

the term for site of the "styloid process" (TEMPORAL BONE). This seems unnecessarily exhaustive because it is reasonable to assume that pain relating to the temporal bone would involve face or head. The simple, less exhaustive formulation TEMPORAL BONE and PAIN would materially have improved recall.

460 Optical or spectral properties of malignant cells which would permit their detection with sufficient efficiency for counting. This request really boils down to "optical and spectral properties of malignant cells". However, in addition to requiring that a neoplasm term should co-occur with an optical property term, the searcher demanded that a "diagnosis" term should also be present. Recall was 76.5% but could have been 100% with a less exhaustive formulation.

In the two previous examples, the exhaustive formulations, although they lost on recall, were at least sensible. The following two examples are not really intelligent:

147 The requester asks for "sodium and potassium ions present in whole blood and erythocytes", but the "ands" are obviously "ors". In other words, he is interested in articles discussing either sodium or potassium in either whole blood or erythrocytes. Inexplicably, the searcher used SODIUM and POTASSIUM and BLOOD and ERYTHROCYTES. As expected, recall was only 25% for this search, although it could have been 100% on

SODIUM
or
POTASSIUM

and

BLOOD
or
ERYTHOCYTES

299 A search relating to "frozen blood platelets" was conducted on:

platelet terms and (BLOOD PRESERVATION or BLOOD BANKS) and (FREEZING or REFRIGERATION or ICE or other refrigeration technic terms)

In other words, the searcher demanded the co-occurrence of two "preservation" terms as well as a term for blood platelets. This was unnecessarily exhaustive and achieved the expected low recall of 25%.

Nonexhaustive formulations

9 The request relates to various aspects of induced hypothermia. This was searched on the single term HYPOTHERMIA, INDUCED. This retrieved 860 citations and predictably obtained 100% recall (25/25). However, the precision ratio was only 30%.

18 In a search on "renal amyloidosis as a complication of tuberculosis", the strategy: amyloid term and tuberculosis term, omitting the requirement for kidney involvement, was responsible for most of the irrelevancy.

211 A specific request for "adverse effects of demethylchlortetracycline on the kidney" was unaccountably searched on the single term DEMETHYLCHLORTETRACYCLINE. The recall estimate is 100%, but the search retrieved 125 citations of which less than 4% are relevant.

#214 The request relates to metabolism and various other specific aspects of mercury radioisotopes. The broadest strategy (level 4) asks only for the coincidence of mercury terms and isotope terms (i.e., it omits the specific aspects requested). The search achieved 90.9% recall (10/11) but only 19.2% of the 273 retrieved citations are relevant. However, the inverse relationship between recall and precision is depressingly brought home by the results for level 5 of the search, which covers all the specific topics requested but only achieved 54.5% recall.

Recall and precision failures due to variations in specificity of the formulation

Only 2.5% of all the recall failures were attributed to the use of a specific search formulation. This does not mean that reduction of searching specificity could not substantially have improved recall in many searches - obviously it could. It merely means that only in the case of 20 missed documents, out of the total of 797 examined, could the blame be put primarily on a search formulation unnecessarily specific in relation to the stated request.

On the other hand, 15.2% of all the precision failures could be attributed to lack of specificity in the search formulation. A "nonspecific" search does not necessarily imply that for the required specific term, A_1, we are substituting the immediately more generic term, A, in the hierarchical tree. Most of the MEDLARS searches are nonspecific in that they substitute for the required specific term, A_1, a term, B_1, from a completely different hierarchy. In other words, instead of asking for A_1 only, the searcher has generalized to say "accept A_1 or B_1." For example, search # 30 relates to "prevalence, incidence and epidemiology of ocular tumors". By using GENETICS, HUMAN as a coordinate with neoplasm terms, the searcher is virtually generalizing to accept "prevalence, incidence, epidemiology, and genetic aspects" of ocular tumors, and we must expect the search to retrieve irrelevant case studies on familial gliomas, retinoblastomas, and other ocular tumors.

In the same way that reduction of exhaustivity in a formulation will tend to improve recall but reduce precision, so reduction in search specificity (if it involves logical generalization) will tend to improve recall but reduce precision. Some examples are given below:

3 A request on "electron microscopy of lung or bronchi" was broadened to LUNG/CYTOLOGY, thus improving recall but inevitably losing precision.

19 In a search on nervous tissue culture, the "tissue culture" was generalized to IN VITRO. This led to about 80% irrelevancy in the search. It seems to be a searching convention that "tissue culture" is generalized to "in vitro studies", with devastating effect on precision. The same thing occurred in search # 91, on skin tissue culture. SKIN and IN VITRO caused the retrieval of about 300 irrelevant items (e.g., on histochemistry, electron microscopy, and biochemistry) out of 777 retrieved.

43 The ease with which it is possible to "explode" on a complete MeSH category, or one of the tree structures, will sometimes lead the searcher into a nonspecific formulation. This requester wanted "immunotherapy of cancer", but an explosion was conducted on the entire C_2, neoplasm, category, thus causing retrieval of irrelevant items on, for example, sarcoidosis.

45 For a search on electrical brain stimulation, the searcher generalized to BRAIN ELECTROPHYSIOLOGY. This led to the retrieval of 533 citations, of which only 17% were relevant, but achieved 83.3% (5/6) recall.

77 The search relates to epidemiology, etiology and genetic aspects of stomach neoplasms. The combinations STOMACH NEOPLASMS and [NEGROES or NEOPLASM STATISTICS or MORTALITY] , which are not specific to the stated request, were responsible for about 60% of the irrelevancy (294 citations were retrieved, of which around 44% were relevant), including many articles on therapy and/or prognosis. Note that the term NEOPLASM STATISTICS has not been used in a strictly epidemiological context; it has also been applied to cover statistics on regression rate, success rate for various therapeutic procedures, and mortality.

101 Again, the ease with which an explosion can be conducted appears to have led the searcher into a very poor search. The request refers to various aspects of personality in relation to choice of medical specialty. The searcher exploded on SPECIALISM and coordinated this set of terms with a group of behavioral terms. Unfortunately SPECIALISM brings out all the terms covering individual medical

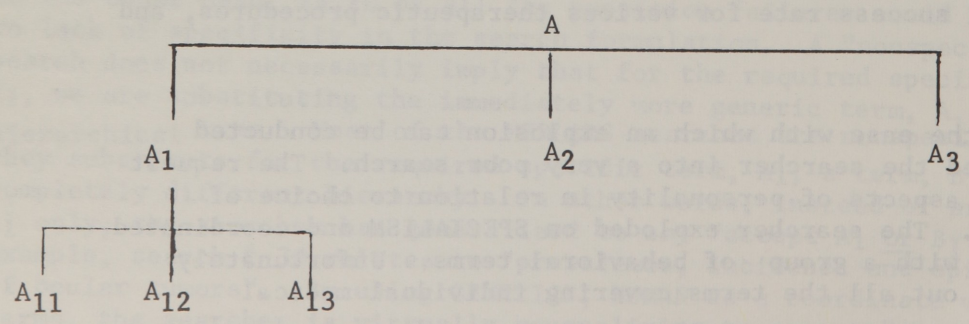
specialties (e.g., PATHOLOGY, PEDIATRICS, PSYCHIATRY, GERIATRICS). Coordinated with the behavioral terms, this retrieved 459 citations (e.g., on personality changes in aging, on doctor-patient relations in pediatrics) of which less than 20 are of any possible relevance.

Although it is terribly dangerous to generalize on the matter of searching strategy, and the correct level of generality to adopt, a detailed study was undertaken to determine if any useful pointers could be derived to assist the searcher in deciding (a) when to broaden a search, (b) the best way to broaden, and (c) what type of search generalization is unwarranted. The details of this study are given in Appendix 5.

In a group of 100 searches examined, 27 were found to include instances of precision failures due to nonspecific search formulations. Each broadened formulation was compared with a formulation that strictly matched the requirements of the request, by matching against the index term profiles of both relevant and irrelevant articles in the precision base to determine (a) how many additional relevant articles were retrieved by the more general strategy, and (b) how many additional irrelevant articles were also brought out. Over the 27 searches, it was found that use of a specific formulation would have avoided 39.6% of the irrelevancy, but would also have lost 17.2% of the relevant items.

The search-by-search analysis of Appendix 5 (a) (nonspecific formulations) and 5 (b) (specific formulations) indicates clearly that in certain of the searches the expansion of the formulation was justified, and recall would have been appreciably lower without it, whereas in other cases the generalization added only irrelevancy. Some general observations on searching strategy can be made on the basis of this detailed analysis:

1. When a requester asks for a specific notion, broadening of the search to the immediate generic term (but not exploding on it) may be justified. Searching on brothers in the hierarchical tree will usually not be justified. Consider the partial hierarchy:



If a request relates to A_1, we would normally expect that articles indexed with the more specific terms A_{11}, A_{12} and A_{13} would be relevant. We might also expect that certain of the articles indexed under A would be relevant. We would not expect that articles indexed under A_2 and A_3, which classes should exclude A_1, would be relevant unless either (a) the request statement is not a good reflection of the actual information need, or (b) the construction of the hierarchy is defective.

This point is well illustrated in search # 166. The requester is interested in a specific tumor (hemangioma) of the small bowel. The searcher used not only the term HEMANGIOMA, but also HEMANGIOPERICYTOMA and HEMANGIOENDOTHELIOMA, and coordinated these terms with terms for "small bowel" (i.e., the search was expanded from A_1 to include also A_2 and A_3). Since hemangiopericytomas and hemangioendotheliomas are not kinds of hemangiomas, there seems little justification for this, although there was justification for the use of HEMANGIOMA, CAVERNOUS (A_{11} in the hierarchy).

On the other hand, the expansion of the search from "articles on hemangioma of the small bowel" to include "general articles on tumors of the small bowel" (i.e., moving from A_1 to search also on the general term A) seems very reasonable because we can expect at least some of these more general articles to include discussion on hemangiomas (although, because of nonspecific indexing, the precise term HEMANGIOMA may not have been applied). In fact, such an elaboration would substantially have improved recall.

A similar situation occurred in search # 221. The requester was interested in medical articles on the Somali Republic. Because the specific term SOMALIA only became available in 1966, the searcher had to use AFRICA, EASTERN for the earlier material. Returning to the hierarchy, this is a move from A_1 to the generic term A. However, the searcher unaccountably exploded on AFRICA, EASTERN. Thus, articles specifically indexed under ETHIOPIA and SUDAN (A_2 and A_3 in the hierarchy), which have no bearing on the topic of the request, were retrieved.

2. Broadening of a search strategy is usually justified when the precise topic of interest to the requester is not adequately covered by appropriate specific terms in the vocabulary. However, when the precise topic of interest is adequately covered by appropriate specific terms, broadening of the search will usually be unjustified. For example, it is very difficult to deal successfully with requests for "ultrastructure" of a particular organ, because no specific terms exactly cover "fine structure". This problem is more acute for the material predating subheadings. For these searches (e.g., # 200 and # 216), the searcher is entirely justified in expanding by the use of "parts" terms (e.g., by exploding on CYTOPLASM) or by the use of "technic" terms (e.g., MICROSCOPY and HISTOLOGICAL TECHNICS). Likewise in search # 177. There appears no good way of precisely expressing the

notion of "premature rupture of the fetal membranes", so that the searcher is certainly justified in broadening to "complications involving the fetal membranes" by searching on

PREGNANCY COMPLICATIONS

FETAL MEMBRANES and or

LABOR COMPLICATIONS

In search # 155, recall suffered because the searcher did not generalize, although it would have been entirely justified. One aspect of the request relates to "analogue models of inert gas exchange". The searcher used specific inert gas terms but might reasonably have broadened to search on GASES since no term exactly covers the inert gases as a group.

In other cases, search elaborations appear entirely unjustified. Search # 212 relates to periodic urine testing as a measure of the adherence of patients to oral tuberculosis therapy (isoniazid and PAS). The requester is only interested in the processes of absorption and excretion, but the searcher generalized to METABOLISM. Similarly, although only urine is of any interest, an explosion was conducted on the entire FLUIDS and SECRETIONS group of terms, which brings out, for example, SPUTUM. Such generalization was responsible for 5/8 of the irrelevancy; it added nothing to recall.

A similar unwarranted generalization occurred in # 268. The area of interest relates to the effect on the orbit of one particular mycosis (covered precisely by the term ASPERGILLOSIS) caused by one particular fungus (covered precisely be the term ASPERGILLUS). The expansion to "all fungal disease of the orbit", by explosions on MYCOSES and FUNGI, appears completely unjustifiable, since it must result in the retrieval of many articles that can have no possible relevance to aspergillosis. Use of the general terms MYCOSES and FUNGI (not exploded) as coordinates, on the other hand, appears reasonable: they could be expected to retrieve more general articles on fungal diseases of the eye that may contain data on aspergillosis.

Search # 557 (not in Appendix 5) contains a further example of unwarranted generalization. The request relates to microwave treatment of food. Since the specific term MICROWAVES has always been available, it is hard to understand why the searcher should generalize to the single term FOOD IRRADIATION. This led to the retrieval of 101 irrelevant items out of 108 retrieved. Such generalization is largely indicative of the searcher's lack of confidence in the indexing.

3. Sometimes a searcher appears to make an unwise choice in deciding which facet of a request to expand on. This was exemplified by

the previously mentioned search on aspergillosis of the orbit. The searcher elaborated in the disease category, and included terms for specific fungal diseases clearly outside the scope of the request. It would have been more sensible to retain the specific terms ASPERGILLUS and ASPERGILLOSIS, but to expand in the anatomy facet by searching on terms relating to the eye in general and to adjacent anatomic structures. The requester stated that aspergillosis of the orbit frequently spreads from one of the sinuses. Inclusion of ASPERGILLOSIS and PARANASAL SINUSES would have improved recall of relevant literature.

Similarly, search # 48 relates to "psychology and rehabilitation of hemiplegics". The searcher held constant the term HEMIPLEGIA, but exploded on the complete F, Psychology and Psychiatry, category. However, many of the F terms are unlikely to relate directly to "psychology of hemiplegics", and will therefore produce irrelevancy. On the other hand, some of the literature relevant to "rehabilitation of hemiplegics" deals more generally with rehabilitation of the "stroke" patient. Expansion in the disease facet to

REHABILITATION and (CEREBRAL HEMORRHAGE or CEREBROVASCULAR DISORDERS)

would have improved recall from 5/13 (38.5%) to 10/13 (76.9%).

4. It is usually disastrous, from the point of view of precision, to explode on two facets of a request simultaneously. That is, given a request for A_1 in relation to B_1, under certain conditions it will be reasonable to hold A_1 constant and expand to B (A_1 and B), or to hold B_1 constant while expanding to A (A and B_1). However, the concurrent expansion of both categories (A and B) will rarely be justified, because it will almost invariably result in extremely low precision.

This is exactly what occurred in search # 174, the subject being "testicular biopsies in infertility and endocrine disease". For the anatomical term TESTIS, disease terms (e.g., TESTICULAR DISEASES, TESTICULAR NEOPLASMS) and "component part" terms (e.g., LEYDIG CELLS, SERTOLI CELLS) were substituted. At the same time, there was a generalization in the technics facet from BIOPSY to general terms such as PATHOLOGY and HISTOLOGY. This led to a precision of only 27.3%.

5. The use of a disease term A, implying some particular site, in coordination with a disease term B, as a way of expressing the site of B, appears to be a searching strategy of doubtful validity, leading inevitably to many false term coordinations and incorrect term relationships. For example, in search # 462, "cerebral amyloidosis" was translated into:

(AMYLOIDOSIS or AMYLOID SUBSTANCE) and terms for CNS diseases

Combinations such as AMYLOIDOSIS and PARAPLEGIA retrieved irrelevant items on, for example, renal amyloidosis in paraplegics.

6. On certain topics, because of the characteristics of the literature and because of indexing conventions, to obtain high recall it is usually necessary to expand the scope of the search. This is true of the subject of "preservation" (see, for example, search # 236 on corneal preservation and search # 238 on heart preservation) which must usually be expanded to "preservation and transplantation" in order to obtain reasonable recall.

Use of "weighted" index terms

The author was surprised to discover, throughout the search analyses, that very little use was made of weighting as a retrieval device, although MEDLARS has a built-in term weighting system in the distinction between print and non-print terms. In less than 5% of all the test searches was use made of print terms to improve the precision of a search.

In the analysis of indexing exhaustivity, the effect of searching only on print terms (i.e., accepting the much lower average exhaustivity level of 2.6 terms per article) was investigated and found, over 88 searches, to lead to a substantial drop in the average recall ratio, from 60% to 44%, with an accompanying rise in the average precision ratio, from 52% to 60%. The effect of searching only on print terms obviously has a drastic effect on recall. With an average of only 2.6 terms per citation, we cannot expect very many to match a two-term coordination in a search formulation.

However, what happens when we retain all the terms assigned to an article but use the print terms as weighted index tags? In the majority of requests, there is a key notion that we would expect to be treated centrally in any relevant article. Consequently, we can reasonably expect that the index term expressing this notion will be a print term. For example, we might reasonably expect that, in the indexing of articles on the action of chloramphenicol (search # 46), the indexer would indicate CHLORAMPHENICOL as being a print term. At least, we would expect that demanding CHLORAMPHENICOL as a print term would retain all the major value articles, although it may lose some minor ones. Moreover, we could reasonably expect that, by weighting this index term, we could screen out much of the irrelevant material, brought out by exhaustive indexing, in which chloramphenicol is mentioned incidentally (e.g., it is used as an incubation medium in a bacterial study).

To test this hypothesis, sixteen searches, based on requests that contained one obviously key notion, were selected for analysis. Brief titles of these searches, indicating the key MeSH term selected for weighting, are given below:

46 Biological effects of CHLORAMPHENICOL.

47 COMPUTER (plus other data processing terms) recognition of cells.

64 Epidemiology of SPINA BIFIDA and ANENCEPHALUS (also MONSTERS).

120 NEUROGLIA cells. Homogolous cells in GANGLIA, AUTONOMIC.

132 Effect of DIGITALIS (and related terms) on gastrointestinal tract.

148 Tubular secretion of CREATINE AND CREATININE.

179 Experimental HYDROCEPHALUS.

208 SYRINGOMYELIA

209 Preparation of radiolabeled FIBRINOGEN (or FIBRIN).

245 Radiation pneumonitis (all lung disease terms demanded to be print).

246 Toxicological, teratological and other aspects of NICKEL.

250 Joint involvement in SARCOIDOSIS.

251 HODGKIN'S DISEASE (and related terms) of animals.

495 Effect of REST on the circulatory system.

509 Effect of HALOTHANE on pulmonary ventilation.

523 HEMOCHROMATOSIS of skin.

Obviously, this is only a small selection of all the test searches from which a single key notion can be isolated, although there are some that it would be difficult to do this with (e.g., in "neurological complications of kidney disease" both notions surely have equal weight). When we average the individual recall estimates for these 16 searches, we arrive at an overall recall estimate of 74.5%. Note that this recall estimate is substantially higher than the recall estimate for the total of 299 searches. This result is simply due to the fact that these searches are, or should be, relatively "simple" searches. Although they may not be based on purely single-term (unifaceted) requests, nevertheless the requests usually involve no more than a relationship between two notions, and one of these notions is clearly an essential notion to the request. Providing that reasonably appropriate index terms exist, as they do in these cases, the system should have no particular problem with this type of search.

The average precision ratio for these 16 searches (derived by averaging the individual ratios) is 48.6%. By analysis, a determination was made of

1. which of the recall base articles for each search would have been retrieved if the "key term" selected was required to be a print term, and

2. which of the precision base (random sample) articles would have been retrieved on the same basis.

It was thus possible to derive, for each search, comparative recall and precision figures based on the strategy of insisting that the key term be a print term. The average recall ratio for the 16 searches dropped, as we would expect, but only to 70.8%, while the average precision ratio increased from 48.6% to 59.7%. Thus, the strategy is a promising one from the point of view of improving precision in this type of search.

This analysis is presented here because it could well be significant in relation to the question of search generality, and how best to "explode" in order to improve recall. We know that broadening the scope of a search is often necessary to obtain an acceptable recall figure. The problem is: how do we broaden to improve recall without having too serious an effect on precision? In some searches, the searcher will throw in every conceivable term to cover a particular aspect of a request (for example, to cover "epidemiology", or "toxicology" or "pulmonary aspects" or "joint involvement"). However, this explosion is carried out on terms that relate to a particular aspect of some major subject that must be present for an article to be of any relevance. That is, they are coordinates of the major term of the search (CHLORAMPHENICOL, SPINA BIFIDA, SARCOIDOSIS are examples from the above searches). Under these conditions, in order to obtain an acceptable recall without too much irrelevancy, it would seem reasonable, while elaborating at great length on the "aspect" terms, to insist that the major term be a print term on any retrieved citation.

One last point in relation to the use of Index Medicus terms as weighted index terms. Not infrequently the searcher will use the term REVIEW in order to retrieve major review articles on a particular subject, as part of a search on some more specific aspect of the topic. For example, in search # 64, although the requester is specifically interested in epidemiology of spina bifida, it would not be unreasonable to expand the search to include review articles on spina bifida, on the grounds that they may well discuss epidemiology, although this aspect is not precisely covered in the indexing. Under these conditions, the term coordinated with REVIEW should always be a print term. This will tend to ensure that REVIEW is in fact related to the topic of the search, and not to some other topic discussed in the article (i.e., it will act as a link to avoid false term coordinations). As an illustration, in one of the test searches (# 68), REVIEW and DECIDUA were coordinated. An irrelevant item retrieved on this coordination does not review the decidua; it is a review of experimental teratology, in which the decidua is merely one of the sites mentioned.

Other causes of searching failures

We have so far discussed the major problems of searching. Tables 6 and 7 also show some "miscellaneous" searching failures, of which a few are worth mentioning. Defective search logic, leading to about 1% of all the precision failures, was found in six of the searches analyzed. This type of error is prone to occur in a highly complex formulation. For example, in search # 13, on blood or urinary steroids in human breast or prostatic neoplasms, the searcher was required to screen out both animal studies and drug studies. Unfortunately, the two negations were placed in an *or* rather than an *and* relationship, thus cancelling each other out, and allowing retrieval of items indexed under animal terms or drug therapy terms. Another type of "logic" failure is the *anding* of a term with a hierarchical tree containing that same term. For example, in search # 479, IMMUNOELECTROPHORESIS was *anded* with an explosion on SERODIAGNOSIS. Since the former term is in the SERODIAGNOSIS tree, the search requirement is reduced to IMMUNOELECTROPHORESIS only. Although such failures do occur, they are rare, and are thus not of too much concern.

About 1.6% of all the recall failures are due to "selective printout". This requires some explanation. Prior to 1967, before ordering a search to be printed, a searcher would receive a notification ("statistical table") from the computer. This table indicated how many citations satisfied the search logic. If the searcher felt that the total was too great, a partial printout could be requested. Under these conditions, sections 6 and 5 of the search, if they existed, would be printed, the residue (up to the total specified by the searcher) being taken from section 4. The sample would be taken as it was read sequentially from the citation tapes, up to the specified total (i.e., the earlier citations would be printed rather than the later). Early in 1967, these procedures were changed. Now there is a ceiling of 500 citations built in to the printing programs. That is, the search printout will be cut off after 500 citations are printed unless the searcher has previously requested that the 500 ceiling be ignored for a particular search. At the same time, a further programming change was made to allow the printing of the more recent citations rather than the earlier ones.

Obviously, in some searches, relevant articles will be among the citations "retrieved" but not printed. This occurred, in fact, in seven of the 302 searches analyzed. Although the selection of the more recent citations is rather more sensible than the selection of the earlier ones, the "sampling" is still no more likely to select relevant citations than irrelevant ones. Again, it would be perfectly feasible to make use of *print* terms as weighting devices to help to ensure that sampled citations are those most likely to be relevant. In fact, if the number of citations satisfying a particular search strategy exceeds 500, it seems unwise merely to accept the built-in cutoff. A retrieval in excess of 500 indicates either (a) that the formulation is defective or imprecise, or (b) that the search topic is a very broad one, upon which considerable literature exists. In the first case, we should reformulate. In the second, we should request a total printout. If partial printouts

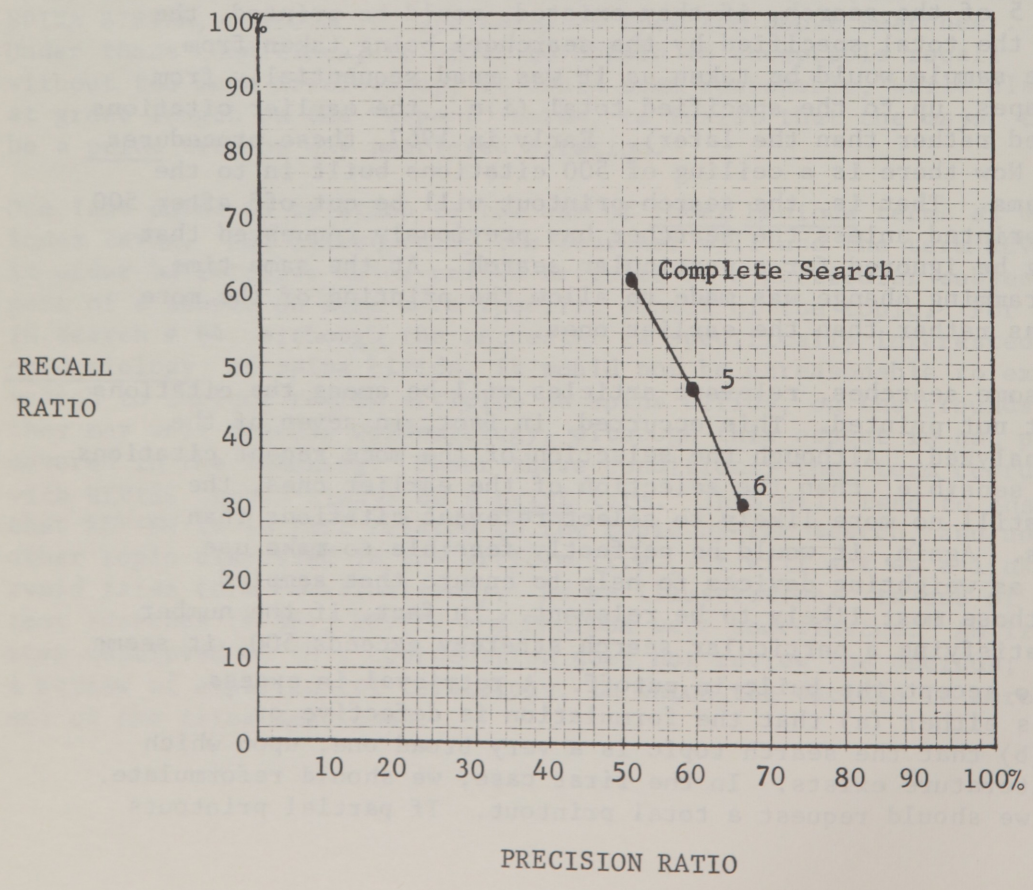
Table 10

Performance figures for 118 searches showing variation with subsorts of increasing specificity (4-5-6)

	Precision ratio	Recall ratio
Complete search (sort 4) (average number of citations = 222)	51.3%	62.7%
Subsort 5 (average number of citations = 124)	59.7%	48.3%
Subsort 6 (average number of citations = 64)	65.7%	32.3%

Figure 7

Three-point performance curve showing variation with subsort of increasing specificity



are to be made, then some ranking method (on the basis of print terms) might usefully be built into the programs.

Joint causes of system failures

Now that both indexing failures and searching failures have been discussed, it is appropriate to re-emphasize the fact that not all failures are attributed to a single cause. Sometimes they have been attributed jointly to two parts of the system. This is particularly true in the case of the relationship between indexing and searching. Occasionally we must say "if the indexer had included X, the document would have been retrieved, but it would also have been retrieved had the searcher used Y". This occurred, for example, in search # 276, on keratinization of the gingiva. The search was conducted on

		GINGIVA
KERATIN	and	GINGIVAL DISEASES
		GINGIVAL HYPERPLASIA
		GINGIVAL HYPERTROPHY

One of the recall base articles, missed by the above strategy, could have been retrieved had the searcher also used KERATIN and GINGIVITIS. On the other hand, the indexer might well have applied the term GINGIVA to this article, as well as GINGIVITIS, because it deals with the effect of powered toothbrushing on gingival inflammation and keratinization. This is the type of failure which, in the analysis, has been jointly attributed to both indexing and searching.

Effect of the 6-5-4 levels on recall and precision figures

As discussed in some detail in Part 1, MEDLARS has the capability of conducting a three-level search of varying specificity in relation to the stated request. Obviously, for those searches in which the three-level strategy is adopted, it is possible to derive separate recall and precision ratios for each of the nesting sets that comprise the search. Within the test corpus, there were 118 searches for which it was possible to do this. There were additional searches with two levels only (section 4 and section 5) but these have not been tabulated as a group. The effects of the three-level searching strategy, averaged over the 118 searches, are shown in Table 10 and Figure 7.

The average size of the complete search was 222 citations, and this particular group of 118 searches achieved a recall ratio of 62.7% at a precision ratio of 51.3%. Clearly, as the search requirements are made more stringent, and fewer documents retrieved, the recall ratio drops and the precision ratio increases. This is shown clearly in Figure 7. At their most specific, the 118 searches averaged 65.7% precision and 32.3% recall. Note that the increments in the recall ratio are greater than the increments in the precision ratio: the move from the most specific searching strategy to the most general almost doubles recall, from 32.3% to 62.7%, while precision drops less the 15%.

Recall and precision failures attributable to the index language

The quality of the index language is probably the most important single factor governing the performance of a retrieval system. Poor searching strategies, and inadequate or inconsistent indexing, can mar the performance of a system, but indexing and searching, however good, cannot compensate for an inadequate index language. In other words, indexers and searchers can perform only as well as the index language allows.

The index language contributed to 10.2% of the MEDLARS recall failures, and 36% of the precision failures. These failures are of two principal types: failures due to lack of specificity in the terms, and failures due to ambiguous or spurious relationships between terms.

Lack of specificity in the index language can cause either recall failures or precision failures. In the present evaluation, it was responsible for 10.2% of all the recall failures and 17.6% of all the precision failures. Although an oversimplification, it is convenient to consider the index language of MEDLARS, or any other retrieval system, as comprising two vocabularies: (1) the controlled vocabulary of terms that indexers must use in describing the content of a document (i.e., the 7000 MeSH terms), and (2) the vocabulary of natural language words and phrases, occurring in documents and requests, that map onto the controlled vocabulary terms. This latter vocabulary we have described as an entry vocabulary. Within MEDLARS, the entry vocabulary is partly built into Medical Subject Headings through the use of references. For example, under CHARCOT-MARIE DISEASE in MeSH we find the instruction see under MUSCULAR ATROPHY. The former term is an entry vocabulary term: it does not uniquely define a class of documents in the system, because the class of "documents on Charcot-Marie Disease" is subsumed under the broader class of "documents on muscular atrophy", and thus has no separate identity. Within the Index Section at NLM is a further entry vocabulary, on 5 x 3" cards, of additional natural language terms that map onto MeSH terms. This entry vocabulary, known as the authority file, consists of about 18,000 entries, of which approximately two thirds relate to drugs and chemicals.

It is worthwhile returning to the earlier discussion on the matter of the entry vocabulary, and to the illustration given in Figure 6. Consider articles on the subject of tetrodotoxin. We decide not to uniquely define this class of documents, but to subsume it under the more generic class "fish toxins", this topic being defined by the joint use of, say, a term ANIMAL TOXINS and a term FISH. Even though we do not uniquely define the class "tetrodotoxin", we must include it in our entry vocabulary, as a reference:

Tetrodotoxin use ANIMAL TOXINS and FISH

We must do this to:

1. indicate that documents on this specific topic have been input to the system, and

2. ensure that all indexers use the same term combination to enter, into the system, articles on this precise topic, and

3. ensure that searchers use the right term combination to retrieve relevant literature.

Thus, although we do not uniquely define the class "tetrodotoxin", we should still be able to retrieve literature on this precise topic, because our entry vocabulary tells us precisely where to look. That is, lack of specificity in the vocabulary will not cause recall failures in this case. However, we cannot retrieve articles on tetrodotoxin alone; we must retrieve the entire class of articles on fish toxins. Thus, lack of specificity will cause precision failures in a search on tetrodotoxin. In other words, if we do not uniquely define a particular class of documents, but still use our entry vocabulary to indicate how this class has been subsumed, we will get precision failures due to lack of specificity in the vocabulary, but not recall failures attributable to this cause. If we omit the notion even from our entry vocabulary, we will get both recall failures and precision failures. Some examples will help to illustrate this point:

6 A search relating to aortic regurgitation had low precision because the 1963 and 1964 material was indexed under the more general term AORTIC VALVE DISEASES.

70 A search on bacterial identification by computer achieved only 33.3% recall and 47.8% precision. There is no specific term for bacterial identification (speciation). The combination of bacteria terms and analysis or automation terms retrieved many articles on bacterial analysis (e.g., chromatographic purification) having nothing to do with species identification. Recall failures in this search can also be attributed to lack of specificity in the index language. "Bacterial identification by computer" is really "numerical taxonomy". There is no term for this, or for "numerical analysis", and nothing in the entry vocabulary to say how it is to be treated. This led to the complete omission of the topic in the indexing of several relevant articles.

84 A search on food preferences and dietary behavior during pregnancy was very unsatisfactory because there was no good way of expressing preferences or behavior. Combinations of pregnancy terms and diet or nutrition terms produced about 170 completely irrelevant articles (e.g., on dietary deficiencies in the Bantu) out of the total of 437 retrieved. The precision ratio was only 22%.

93 The search was for articles on hypophosphatasia. But a provisional term for this concept became available only on 7/19/66. It is extremely difficult to retrieve literature input prior to this date. The searcher tried

METABOLISM, INBORN ERRORS	and	BLOOD ALKALINE PHOSPHATASE ALKALINE PHOSPHATASE PHOSPHATASES

but would have needed to search on the single term BLOOD ALKALINE PHOSPHATASE (with over 800 postings) to obtain a satisfactory recall.

102 A search on the use of Fourier series in hemodynamic analysis achieved only 54.5% recall and 8% precision; 99 citations were retrieved. There is no specific term for "Fourier analysis", and nothing in the entry vocabulary to say how this notion is to be subsumed. Consequently, the concept was omitted in the indexing of several pertinent articles. Moreover, the searcher was forced into very general combinations (e.g., hemodynamics terms and MATHEMATICS or MODELS) which caused considerable irrelevancy.

160 This search, on nephrogenic diabetes insipidus, clearly illustrates the importance of an adequate entry vocabulary. In an attempt to restrict the search to nephrogenic diabetes insipidus, the searcher coordinated DIABETES INSIPIDUS with kidney or kidney disease terms. This resulted in the low recall ratio of 1/9 (11.1%) because the topic is generally indexed only under DIABETES INSIPIDUS, with no co-occurring kidney term. However, the searcher had no way of knowing this because there is nothing in the entry vocabulary to say how this notion is to be indexed.

177 This search on "premature rupture of the fetal membranes" was a total failure. There is nothing in the entry vocabulary to say how this notion is to be treated. Consequently indexers have omitted the concept in the indexing of several pertinent articles. Where the indexer has attempted to cover it, there has been no consistency in the terms used. Highly relevant articles were found under all of the following terms, or term combinations:

LABOR, PREMATURE and PREGNANCY COMPLICATIONS
LABOR, PREMATURE and FETAL MEMBRANES
LABOR COMPLICATIONS and FETAL MEMBRANES
PREGNANCY COMPLICATIONS and FETAL MEMBRANES
RUPTURE, SPONTANEOUS and FETAL MEMBRANES

180 Indirect pulp capping is subsumed under DENTAL PULP CAPPING. Such lack of specificity should not affect recall, but will affect precision. The more general term retrieved 71 citations of which about 45% were relevant to the indirect process.

181 Poor results were obtained in a search on asymptomatic proteinurias. There is no good way of expressing "asymptomatic,"

so only 17.4% precision was achieved. The searcher used ALBUMINURIA and PROTEINURIA, and attempted to restrict the search to asymptomatic proteinurias by negating kidney disease terms. Unfortunately, this has no effect in keeping out proteinurias in other disease. Moreover, it also screens out some relevant articles that legitimately have a kidney disease term assigned to them.

190 A search on "deiodination of thyroxine" was carried out on THYROXINE and IODINE or IODINE ISOTOPES, because there is no specific term for "deiodination". This led to the retrieval of 628 citations, of which only 36.8% were of any relevance. Despite the large retrieval, recall (85.7%) was less than maximum, because lack of an entry vocabulary term for "deiodination" has led to the omission of this notion in the indexing of relevant articles.

207 The request was for phosphorylase deficiency myopathy (McArdle's disease). Because there is no entry vocabulary term, the searcher was forced to try many different combinations of muscular disease terms and such terms as GLYCOGEN, GLYCOGENOSIS, and PHOSPHOTRANSFERASES. These retrieved 200 citations and achieved 100% (8/8) recall, but only 44.8% precision. In actual fact, all of the recall base articles could have been retrieved on the limited set of combinations:

GLYCOGEN		MUSCULAR DISEASES		PHOSPHOTRANSFERASES
or	and		and	or
GLYCOGENOSIS				PHOSPHORYLASE KINASE

217 A search on the styloid process in facial and head pains was conducted on TEMPORAL BONE and pain terms, because there is no term for "styloid process" or even "process" in general. There was almost 90% irrelevancy in the search. From analysis, it appears that indexers have tended to use the term ABNORMALITIES for the notion of "process". The searcher does not know this, and there is nothing in the entry vocabulary to so inform her. Three of the four known relevant articles could have been retrieved, probably with 100% precision, on the combination:

ABNORMALITIES and TEMPORAL BONE and NEURALGIA

235 The specific disease entity "colitis cystica profunda" appears nowhere in the vocabulary. Nonspecific combinations, such as COLITIS and CYSTS, COLON and CYSTS, retrieved 75 citations and achieved 66.7% recall but only 8.3% precision.

242 The fact that "masers" must be translated into MICROWAVES led to the retrieval of 304 citations of which only 20% were relevant.

453 A search on gallbladder perforation achieved only 33% recall at 40% precision because the notion does not appear explicity anywhere in the vocabulary.

#455 This search exemplifies the general inadequacy of the vocabulary in the behavioral sciences. To express perceptual completion phenomena, the searcher was forced into very general combinations (e.g., VISION and ILLUSIONS) which, although they retrieved 173 citations, achieved only 16.7% recall at 10.5% precision.

467 A search on the effect of high frequency radio waves (Diapulse) on wound healing operated at only 10% precision. There is no specific term for "diapulse", and MeSH refers from Short-wave therapy to DIATHERMY, which is misleading since not all short-wave therapy is diatherm. In fact, "diapulse" is athermic, but some material on the process has been indexed under DIATHERMY, while other material is under MICROWAVES.

485 A search on the Hallervorden-Spatz syndrome was unsuccessful because the specific term only became available, as a provisional heading, on 2/13/65. There is nothing in the entry vocabulary to say how the earlier material was indexed. The searcher tried BRAIN DISEASES and GLOBUS PALLIDUS and SUBSTANTIA NIGRA, but much of the earlier material is indexed only under GLOBUS PALLIDUS.

506 A search on finger tip amputations had to retrieve articles on whole digit replacement, because it is possible to express "fingers" but not "fingertips".

511 "Left ventricular bypass" is too specific for the vocabulary; more general combinations such as HEART, ARTIFICIAL and HEART VENTRICLE retrieved 500 citations, of which only 38% were relevant, and still attained less than 60% recall.

530 Again, nothing in the entry vocabulary to say how "spongy degeneration" (of the white matter) is to be indexed or searched for. The search achieved only 4/11 (36.4%) recall. There is no consistency in how this topic has been indexed. CNS - related terms assigned to relevant articles are as follows:

BRAIN EDEMA and CONVULSIONS
BRAIN DISEASES
BRAIN DISEASES and NERVE DEGENERATION
CENTRAL NERVOUS SYSTEM DISEASES and DEMYELINATION
BRAIN DISEASES and CEREBRAL CORTEX

545 To retrieve articles on "pseudotumor formation in hemophilia and Christmas Disease", because there is no specific term for "pseudotumor", the searcher had to coordinate HEMOPHILIA with a long list of bone disease terms. This achieved high recall (83.3%) but low precision: only 13% of the 66 articles retrieved are of any relevance. Articles on pseudotumors of hemophilia have been variously indexed under:

HEMOPHILIA and BONE DISEASES
HEMOPHILIA and JOINT DISEASES
HEMOPHILIA and CALCANEOUS

551 Although the specific entity Asherman syndrome (intrauterine synechiae) appears nowhere in the vocabulary, the searcher was able to obtain 83.3% recall at 57.1% precision by coordinating UTERUS and ADHESIONS. It is interesting to compare this search with # 545 which achieved a comparable recall but at a much lower precision. The reason is that we can get very close to the notion of "intrauterine synechiae" by the combination UTERUS and ADHESIONS. We cannot get anywhere close to "pseudotumor" by any combination in the vocabulary.

603 "Acute cecitis" must be translated into either CECAL DISEASES, or CECUM and INFLAMMATION. These retrieved 121 citations, of which but a handful were relevant, and achieved only 33.3% recall.

The above examples illustrate overall index language deficiencies in MEDLARS, and suggest areas (e.g., the behavioral sciences) in which the system is particularly weak. To pinpoint more precisely the subject areas in which the vocabulary is suspect, a breakdown by subject field was made for all the searches in which were found recall and/or precision failures due to lack of specificity in the index language. The results are presented in Table 11.

It can be seen from this table that over one third of the searches falling in the general area of the behavioral sciences are marred because of lack of specificity in the vocabulary, while one third (2/6) of the searches relating to public health are similarly affected. In the area of "technics", 27.6% of all the searches are affected by lack of specificity. On the whole, the language in this area is reasonably unambiguous. However, performance will depend upon the availability of specific terms. As already noted, there is no term covering high

Table 11

Recall and/or precision failures due to lack of appropriate specific terms in the index language. 302 searches were examined, and in 71 of these were found recall and/or precision failures due to lack of specificity in the vocabulary. Breakdown by subject field of request.

Subject field	Number of searches in which failures due to lack of specificity occurred	Percentage of total searches involving lack of specificity (71)	Percentage of total searches for requests in this subject field
PRECLINICAL SCIENCES*	15	21.1%	17.6%
DISEASE	26	36.6%	23.6%
TECHNICS	16	22.5%	27.6%
DRUG/ BIOLOGY	5	7.0%	18.5%
DRUG/ DISEASE	3	4.2%	21.4%
BEHAVIORAL SCIENCE	6	8.5%	35.3%
PHYSICS/ BIOLOGY	3	4.2%	25.0%
PUBLIC HEALTH	2	2.8%	33.3%

* The subject categories are not mutually exclusive. Certain multi-faceted requests were assigned to more than one category.

frequency radio therapy (diapulse) so that search # 467 could operate at only 10.5% precision.

A quarter of the PHYSICS/BIOLOGY searches are affected by lack of specificity in the vocabulary (it is difficult to distinguish various types of radiation; e.g., ionizing from non-ionizing). One surprising fact emerges from Table 11. Although MEDLARS is often accused of being largely clinically oriented, 23.6% of the searches falling into the DISEASE category are affected by lack of specificity, whereas only 17.6% of the PRECLINICAL SCIENCES searches are similarly affected. In other words, for the types of requests put to MEDLARS in the preclinical area, the vocabulary is shown to be reasonably adequate. The performance in the disease area will depend entirely upon whether or not terms for specific disease entities are available in either MeSH or the entry vocabulary. If we have no term for "colitis cystica profunda" we can hardly expect to achieve a satisfactory result in a search on this topic. On the whole, the search analyses have shown the MEDLARS vocabulary to be unexpectedly weak in the clinical area. Not only does it fail to express precisely a significant proportion of the pathological conditions occurring in requests, some of which are not particularly obscure (e.g., perforation of the gallbladder), but it is also deficient in its ability to express various characteristics of a disease. For example, we cannot indicate extent of pathological involvement (# 567, diffuse lesions of the lung, and # 570, solitary pulmonary nodule). Nor can we readily distinguish: acute from chronic; versions of a disease according to etiology (e.g., bacterial from non-bacterial asthma); symptomatic from asymptomatic; co-existent, unrelated conditions from true sequelae; or the situation of one disease simulating ("masquerading as") another.

Again from the search analyses, the vocabulary appears weak in areas that impinge upon medicine. For example, in search # 102, the terms MATHEMATICS, MODELS and COMPUTERS are as close as the searcher was able to get to the topic of "Fourier analysis". Similarly, in search # 242, MICROWAVES was the only term available to express "masers".

Before leaving the matter of specificity in the vocabulary, it is worthwhile mentioning, or re-emphasising, the following:

1. There is a difference between failures due to lack of specificity in the vocabulary and failures due to lack of specificity in searching. In the former case, there is no specific term available so the searcher is obliged to use more general terms. In the second case, the searcher broadens the search even though more specific terms, of varying degrees of appropriateness to the request, exist in the vocabulary. As an example, we can consider the search on Somalia (# 221). Lack of a specific term before 1966 made it necessary for the searcher to include AFRICA, EASTERN in the formulation (lack of specificity in the index language), but the searcher went beyond the deficiency in the vocabulary by exploding on AFRICA, EASTERN, and thus bringing in articles indexed under ETHIOPIA and SUDAN (lack of specificity in searching). In some searches, although there was no specific MeSH term to cover a request topic (e.g., searches

200 and # 216, relating to ultrastructure), the author felt that the searcher generalized more than was necessary. In such cases, some of the failures were attributed to lack of specificity in the vocabulary, others to lack of specificity in searching.

2. To correct precision failures due to lack of specificity, requires that terms or term combinations that uniquely define the notion not presently covered specifically, be introduced into the vocabulary. To correct recall failures, we do not need a unique designation, but we must include the notion in our entry vocabulary.

3. The evaluation has shown the MEDLARS entry vocabulary to be very inadequate. Recall failures in the test searches could have been reduced by 10% if a satisfactory entry vocabulary had been available. Lack of an adequate entry vocabulary can lead to:

1. Indexer omissions, or lack of exhaustivity of indexing (the indexer does not know how to index a particular notion so leaves it out).

2. Indexing inconsistencies.

3. Recall failures.

4. Precision failures (the searcher does not know how a particular notion has been treated, and is thus forced to use every possible term combination). Moreover, the fact that a term appears in an entry vocabulary indicates that literature on the topic exists in the system. Without such an entry, we have no assurance that MEDLARS even contains any articles on, say, some obscure syndrome. Consequently, we may willingly accept a negative result from the system when such a result is incorrect.

The value of the entry vocabulary is well illustrated by search # 532 on "irradiation of mammalian oocytes". The searcher relied on OVUM and irradiation terms. The authority file contains an entry, dated July 31, 1966, which instructs

Oocytes use OVUM.

But articles indexed before this date were indexed with no consistency (some were indexed under GERM CELLS), while this precise notion was omitted in various other articles (indexed only under CELL DIVISION).

4. Searching difficulties are caused by the fact that the vocabulary has been developed without consistency as to levels of specificity (degree of pre-coordination). Thus we have a specific term VAGOTOMY, for example, but we cannot express "pyloroplasty" (see search # 474) except by PYLORIC STENOSIS and SURGERY, OPERATIVE, or by PYLORUS/SURGERY.

Consequently, although we can say VAGOTOMY/ADVERSE EFFECTS, we have no precise way of expressing "adverse effects of pyloroplasty"

5. The MEDLARS index language is gradually becoming more specific. Not all of the failures attributed to lack of specificity indicate current inadequacies. That is, in some cases terminological changes have been made, but the searcher is still required to use nonspecific terms to retrieve material indexed before the specific terms were introduced. To discover what proportion of the failures, attributed to lack of specificity, represent terminological changes since rectified, and what proportion represent terminological deficiencies still existing, a special analysis of a sample of 100 searches was conducted.

In this group of 100 searches, 24 contained recall and/or precision failures due to lack of specificity in the vocabulary. There were 25 separate terms involved, and 13 (52%) of these deficiencies no longer exist in MEDLARS (i.e., appropriate specific terms are now available).

Note that the introduction of subheadings, in 1966, markedly increased the specificity of the vocabulary. It is now possible to express various notions (e.g., "epidemiology" and "etiology") which were not adequately covered in the vocabulary before the subheadings were introduced.

Failures due to false coordinations and incorrect term relationships

Ambiguous and spurious relationships between terms accounted for approximately 18% of all the precision failures. In one sense, all terms assigned in the indexing of a particular article must be considered related in some way, even if it is only a proximity relationship (i.e., the terms are common to a particular index term profile). However, consider a search involving a simple two-term logical product relation, A in relation to B. Although all the articles retrieved by this coordination should be indexed under the term A and also under the term B, some of these articles may be irrelevant because the term A and the term B are not directly related in the article (a false coordination), while others may be irrelevant because, while A is related to B, the terms are not related in the way that the requester wants them (an incorrect term relationship).

To clarify the distinction between false coordinations and incorrect term relationships, consider search #61, relating to phosphate excretion. A term combinations used in this search was PHOSPHATES and URINE. One of the articles retrieved by this combination discusses urinary excretion of Toxogonin, which is mentioned as being an antidote to alkyl-phosphate poisoning. Hence, the term URINE is not directly related to the term PHOSPHATE (i.e., it is a false coordination). The same terms retrieved a second article, not on excretion of phosphates,

but on a phosphate precipitation method of determining magnesium in urine. This is an incorrect term relationship: URINE is related to PHOSPHATE, but not in the way that the requester wants these terms related.

False coordinations were responsible for 11.3% of the precision failures, incorrect term relationships for 6.8%. Some further examples are given below. False coordinations:

71 In a search on mongolism occurring with leukemia, the combination LEUKEMIA and MONGOLISM retrieved a number of articles in which the two terms refer to different patients (e.g., general articles on sex-chromatin abnormalities). This type of failure is always likely to occur in MEDLARS when two disease terms are coordinated. In # 499, on the neuropathy of multiple myeloma, the coordination of MULTIPLE MYELOMA and neurological disease terms caused about 40% irrelevancy, while search # 103, on ventricular septal defect in association with mitral stenosis or mitral insufficiency, achieved only 26% precision because in most of the retrieved articles the two terms are not related.

72 SOMATOTROPIN and REVIEW retrieved a review, not of somatotropin, but of insulin. This type of false coordination is easily avoided by the use of print terms in searching.

96 LUNG and LYMPH NODES retrieved articles that do not deal with pulmonary lymphatics (the two are discussed separately).

492 The requester is interested in combinations of various topical medicinal agents (resorcinol, sulfur compounds, allantoin and hexachlorophene). Combinations of two drug terms retrieved a great many articles in which the two terms are essentially unrelated.

Incorrect term relationships

39 The search relates to cases of prolonged amenorrhea or infertility following discontinuance of oral contraceptives. But about a third of the articles retrieved deal with therapeutic use of contraceptive agents in the treatment of menstruation disorders, and not with side effects. This search illustrates the principal type of relational indicator needed by MEDLARS, namely an indicator of sequence or cause-effect relationship. The same type of failure is likely to occur in searches on radiation and drug effects, because it is sometimes difficult to distinguish therapy from adverse effects. For example, in search # 95, on the effect of radiation on hair growth, HAIR REMOVAL and RADIOTHERAPY retrieved articles on therapeutic use of irradiation (e.g., in alopecia mucinosa) as well as articles on radiation damage to hair.

67 In a search on lipids in annelids, the combinations CEPHALINS and LEECHES and CHOLESTEROL and NEMATODA retrieved a number of articles not on lipids of worms, but on the effect of lipids on worms (e.g., effect on leech muscle preparation).

73 In a search on bovine leukosis, CATTLE and LEUKEMIA retrieved articles not on cattle leukosis but on the reaction of sera from human leukemia patients, or the reaction of mouse leukemia viruses, with bovine cell cultures.

159 In a search on oral hypoglycemic agents in the therapy of juvenile diabetes, DIABETES MELLITUS and TOLBUTAMIDE retrieved a number of articles not on therapy, but on the "tolbutamide tolerance test".

165 The search relates to morphological changes resulting from muscular exercise, including exercise-induced hypertrophy. HEART ENLARGEMENT and EXERTION retrieved a number of articles on the effect of exercise on subjects with heart disease and ventricular enlargement (rather than heart enlargement following exercise).

251 The coordination of HODGKIN'S DISEASE and animal terms retrieved a number of articles not on Hodgkin's Disease of animals, but on experiments with human Hodgkin's Disease, using laboratory animals. This type of failure can now be avoided by use of the subheading VETERINARY.

Table 12 and Table 13 present breakdowns by subject field of the failures due to false coordinations and incorrect term relationships. It can readily be seen that this type of failure is much more prone to occur in some subject areas than in others. No less than 58.3% of the PHYSICS/BIOLOGY searches are marred by this type of failure. This is due to the problem of distinguishing, at least before the arrival of subheadings, radiation injury from radiation therapy. False coordinations affect 43.6%, and incorrect term relationships 30%, of all the 110 searches falling into the DISEASE group. As previously mentioned, when we coordinate two disease terms, we are likely to retrieve articles in which the terms refer to different patients, or, if they refer to the same patient, the relationships between the diseases is not the one required (B causing A rather than A causing B).

Searches relating to drugs are also likely to lead to false coordinations (two drug terms are not related, or a drug term is not related to the specified disease term) and incorrect term relationships (the drug B is used therapeutically in a case of A, whereas the requester wanted cases of A resulting from the use of drug B).

In the literature of documentation, the solution generally offered to the problem of false coordination is the link, while the solution generally offered to the problem of incorrect relationships is the role indicator. However, in the search analyses it was repeatedly discovered that both types of problem could now frequently be avoided by the use of subheadings.

Table 12

Precision failures due to false term coordinations. 302 searches were examined, and 118 found in which this type of failure occurred. Breakdown by subject field of request.

Subject field	Number of searches involving false coordinations	Percentage of total searches involving false coordinations	Percentage of searches for requests in this subject field
PRECLINICAL SCIENCES	28	23.7%	32.9%
DISEASE	48	40.7%	43.6%
TECHNICS	18	15.3%	31.0%
DRUG/BIOLOGY	7	5.9%	25.9%
DRUG/DISEASE	5	4.2%	35.7%
BEHAVIORAL SCIENCE	5	4.2%	29.4%
PHYSICS/ BIOLOGY	7	5.9%	58.3%

Table 13

Precision failures due to incorrect term relationships. 302 searches were examined, and 93 found in which this type of failure occurred. Breakdown by subject field of request.

Subject field	Number of searches involving incorrect relationships	Percentage of total searches involving incorrect relationships	Percentage of searches for requests in this subject field
PRECLINICAL SCIENCES	22	23.7%	25.9%
DISEASE	33	35.5%	30.0%
TECHNICS	13	14.0%	22.4%
DRUG/ BIOLOGY	9	9.7%	33.3%
DRUG/ DISEASE	5	5.4%	35.7%
PHYSICS/ BIOLOGY	7	7.5%	58.3%
BEHAVIORAL SCIENCES	3	3.2%	17.6%
PUBLIC HEALTH	1	1.1%	16.7%

Moreover, failures of this type were found to be virtually nonexistent when indexer and searcher had both made correct use of subheadings. This led the writer to undertake an investigation to determine just what proportion of all the failures of this type could be corrected by the use of subheadings (either existing subheadings or subheadings that could readily be devised). The results are presented in Appendix 6.

In 45 searches examined, 20 examples of false coordinations, and 22 examples of incorrect term relationships, were encountered. A total of 16 (80%) of the false coordinations could have been prevented by subheadings, 12 of these by existing subheadings and 4 by suggested new subheadings. A total of 20 (90%) of the incorrect term relationships could be prevented by subheadings, 14 by existing subheadings and 6 by suggested new subheadings.

Full details of this analysis are given in Appendix 6, and need not be repeated here. It is sufficient to say that subheadings, properly used, are capable of solving 80-90% of the precision failures attributable to false coordinations and incorrect term relationships. The subheadings ADVERSE EFFECTS and THERAPEUTIC USE serve to distinguish articles on therapy from articles on side effects. ETIOLOGY is another subheading useful in obviating the sequential or cause-effect type of problem. The subheading COMPLICATIONS tends to tie two disease terms together and thus avoid some of the false coordinations that occur when we and these terms. That is, in an article indexed disease A/COMPLICATIONS and also disease B/COMPLICATIONS there is a high probability that both conditions co-exist in the same patient. However, the subheading COMPLICATIONS does not solve the sequential problem. Does condition A lead to condition B, or does B lead to A? It would be necessary to introduce a new subheading SEQUELAE to cope with this type of situation.

More freely available subheadings would tend to reduce problems stemming from variations in the specificity (by pre-coordination) of the vocabulary. We can now say BLOOD PRESERVATION, but we can only express "plasma preservation" by the coordination of BLOOD PRESERVATION and PLASMA, which leads to false coordinations. A generally applicable subheading PRESERVATION, in place of the pre-coordinations that exist in parts of the vocabulary, would solve this type of problem.

Of particular value within MEDLARS are paired subheadings. That is, subheadings that tend to tie two terms together and at the same time indicate the relationship between these terms. Such subheading pairs act simultaneously as links and as roles. They function in much the same way as the paired role indicators introduced by Western Reserve University (property given and property given for) and the Engineers Joint Council (causative agent, thing affected). For example the coordination of TOLBUTAMIDE and DIABETES MELLITUS, in a search on therapeutic use of oral hypoglycemic agents in diabetes, will retrieve irrelevant articles on the "tolbutamide tolerance test". But the coordination TOLBUTAMIDE/THERAPEUTIC USE and DIABETES MELLITUS/DRUG THERAPY not only tends to tie the two terms together, but also shows their relationship.

We also feel that additional subheadings might well solve the problem of lack of specificity in the area of disease. That is, attached to disease terms, they could be most useful in indicating "characteristics" of the disease. The search analyses have indicated that the following (by no means a complete list) would all be useful in increasing specificity, and avoiding false coordinations and incorrect term relationships, in the disease area:

CHRONIC

ACUTE

UNKNOWN ETIOLOGY

BACTERIAL (some diseases may or may not be bacterial or viral in origin)

VIRAL

SIMULATED (for the case of one condition masquerading as another)

DIFFUSE
LOCALIZED (for extent of involvement)

DIFFERENTIAL DIAGNOSIS (a particular disease is not discussed in itself but only in the differential diagnosis of some other disease)

EXPERIMENTAL

It must be emphasized that, in the above list and in the additional suggestions contained in Appendix 6, the author is not putting forward what he considers to be a final set of subheadings that should be incorporated into the system. These are merely illustrative. They are certainly not complete, nor are they necessarily the best subheadings to solve the various problems encountered. They do, however, show clearly that additional subheadings, carefully selected on the basis of an analysis of relationships demanded in MEDLARS requests, can obviate many of the failures presently attributed to (1) lack of specificity in the vocabulary, (2) false coordinations, and (3) incorrect term relationships.

Failures due to defects in the hierarchical (tree) structure

Seeming defects or anomalies in the hierarchical structure of the vocabulary were partly responsible for precision failures in five of the searches, but contributed to only 0.3% of all the precision failures. The following are examples:

20 Growth, regeneration and degeneration in the nervous system. An explosion on C_{10}, nervous system diseases, brings out the term PAIN. Coordinated with the other search terms, this led to some peculiar results. For example, PAIN and WOUND HEALING retrieved an article on hemorrhoids.

481 An explosion on $E_{3.26}$, ANESTHESIA, brings out the term INTUBATION, INTRATRACHEAL, which is not exclusively related to anesthesia. In this anesthesia search, it was responsible for the retrieval of an article on intratracheal administration of polonium in a toxicity study.

General observations on the MEDLARS index language

1. There are certain types of requests being made to MEDLARS which are attempted, but with which the vocabulary is completely unable to cope. Obvious examples are search # 479, which covers complex inter-relationships of immunology ("the relationship described by the action of varying quantities of viral antigen with a constant amount of homologous antiserum, or by the reaction of varying quantities of viral antiserum with a constant amount of homologous antigen") and search # 503, which seeks articles on osteomyelitis of unknown etiology.

2. Even with the tree structures, the vocabulary is not as helpful as it could be to indexers and searchers. It is difficult sometimes to think of all terms that are possibly related to a request. Further relationships, built into the hierarchical displays, could be of great assistance to the searcher, and might well help to reduce those recall failures attributed to the searcher not covering all reasonable approaches to retrieval.

3. The author feels strongly that the methods presently used to update the MEDLARS vocabulary are not optimally responsive to the requirements of the demand search function. Heavy reliance is placed on committees of subject specialists to review terminology in particular areas. The use of such committees tends, of course, to ensure that MeSH reflects current medical terminology. This may be highly desirable for the published bibliography, Index Medicus, but is not necessarily the principal requirement for vocabulary development in a retrospective search system based on the coordination of terms at the time of searching.

A vocabulary tends to be most responsive when it has a high degree of literary warrant. In other words, the most valuable raw materials for vocabulary development are incoming articles and, crucial, requests being made to the system. Yet these are the very materials that appear most neglected in the development of the MEDLARS index language.

Within the present evaluation program, requests have been systematically analyzed from the point of view of the capability of the vocabulary to cope with them, but this is not done as part of the regular operations of the system. Although a form (Request for Medical Subject Heading Change) is available to record suggestions of indexers and searchers, very little use appears to be made of this. In other words, there are no routine, established procedures whereby indexers and searchers are required to notify the MeSH group whenever they discover either (a) an article on a topic that cannot adequately be covered in indexing, or (b) a search which cannot be conducted, or can be conducted very imperfectly, because of vocabulary inadequacies.

Consequently, no adequate entry vocabulary has been developed, indexing omissions are caused by the fact that no appropriate terms are available and indexing inconsistencies also occur. This leads to the failure of certain searches (for example, that on "premature rupture of the fetal membranes" and the one on "gallbladder perforation") that should be well within the capabilities of the sytem. Moreover, since searchers do not automatically inform the MeSH group of such topics, upon which they find it difficult to conduct an adequate search, these problems are perpetuated in the system.

4. Although subheadings were apparently introduced primarily to facilitate effective use of the published bibliographies, these subheadings, as the analyses have shown, are of great potential value in reducing precision failures due to false coordinations and incorrect term relationships. The subheadings also afford an economical means of substantially increasing the specificity of the index language.

For example, the notion of "preservation" is applicable to many of the anatomical terms in the vocabulary. However, it would obviously be uneconomical to add to MeSH a substantial number of precoordinated terms incorporating "preservation". In actual fact, only BLOOD PRESERVATION and TISSUE PRESERVATION presently exist. However, the addition of a freely available PRESERVATION subheading adds greatly to the specificity of the vocabulary, does not increase the size of MeSH, and, by linking notions together in indexing, avoids the false coordinations that occur, for example, when we coordinate BLOOD PRESERVATION and PLASMA in an attempt to express "plasma preservation".

5. So far the searchers appear to have been remarkably successful in compensating for vocabulary changes made in the period 1964-1967. Very few recall failures could be attributed directly to the fact that the searcher did not make use of all the terms necessary to retrieve literature on a particular topic, because of MeSH changes over the years. However, as more changes are made to the vocabulary, to make it more responsive to the demands placed upon it, searching is likely to become more and more complex. Moreover, extensive vocabulary changes tend to have a drastic effect on the economics of the search process. It is time-consuming to establish that to conduct a comprehensive search on the epidemiology of a particular disease, we must use a certain set

of terms for the 1964 material, others for 1965, and add subheadings for the 1966 and subsequent material. Although changes are obviously justified if the vocabulary is shown to be deficient in particular areas, the author fears that these changes are making the searching process unnecessarily esoteric. A possible solution, worth investigating, is the use of automatic term substitution by computer. For example, in conducting a search on "circadian rhythms", the searcher should not be required to remember (or to establish in some authority list) that the term PERIODICITY must be put down to retrieve articles prior to the introduction of the specific CIRCADIAN RHYTHM. Whenever the term CIRCADIAN RHYTHM appears in a search formulation, the term PERIODICITY (with the appropriate date restriction) could automatically be added by computer program. Besides improving the economics of searching, such a procedure would avoid the type of failure encountered, for example, in search # 307, on mouse wasting syndrome after thymectomy, which was conducted on MICE and THYMECTOMY and HOMOLOGOUS WASTING DISEASE. Unfortunately, THYMECTOMY came into use only in 11/64, and THYMUS GLAND is necessary to retrieve the earlier material. This search missed 66.7% of the relevant literature as a result.

The relationship between indexing, index language, and searching

Although not something that can be proved in any statistical sense, the author feels that some of the problems discussed in relation to indexing, searching and index language, stem from the fact that these functions tend to be compartmentalized at NLM. The Index Section, the Search Section and the MeSH group, although they may meet periodically to discuss various problems, are self-contained units that appear to operate largely independently. The prime goal of indexing is, presumably, to describe documents in such a way that they may later be retrieved in response to requests for which they are likely to contain relevant data. However, the great majority of the indexers do not prepare searching strategies, and no mechanism exists to keep indexers informed on the types of requests being put to the retrospective search system. Likewise, the analyses have shown that searchers are not fully aware of indexing protocols. For example, searches on tissue culture are frequently broadened to IN VITRO, although the indexers claim that they always use the specific term TISSUE CULTURE for these studies. A search on "premature rupture of the fetal membranes" (# 177) was conducted on RUPTURE and RUPTURE, SPONTANEOUS, whereas most of the relevant literature is indexed under

FETAL MEMBRANES and LABOR COMPLICATIONS
or
PREGNANCY COMPLICATIONS

and the indexers claim that the "rupture" terms are inappropriate to this search since they refer to traumatic rupture, whereas "premature rupture" is a normal physiological process. Again, indexers appear to be using the term ABNORMALITIES for "process", but the analyst who

prepared the formulation for search #217 does not seem to know this. Likewise (search # 160, kidney and kidney disease terms were coordinated with DIABETES INSIPIDUS to express "nephrogenic diabetes", but it has not been indexing policy to use kidney terms in this case.

From the observations of the author, during the conduct of the test, the relationship between indexing and searching is not one of full cooperation towards a mutual goal. Indexers claim that searchers are "not using the correct terms"; the counter-claim of searchers is that they must "compensate for indexing inadequacies". The further separation of Medical Subject Headings from both the indexing and the searching functions, which has resulted in the failure to base vocabulary development on inputs from indexers and searchers, is felt to be no more healthy than the divorce of indexing and searching.

Recall and precision failures attributable to the area of user-system interaction

Defective user-system interaction contributed to 25% of the recall failures and 16.6% of the precision failures in the present evaluation. Note that a few of the precision failures are judged "inexplicable". These are cases in which a retrieved article was judged of no value although it appears within the scope of the stated request, and the author has not been able to determine exactly why the requester found it irrelevant.

A recall failure due to defective interaction implies that the stated request is more specific than the actual area of information need (Figure 8). Articles of value to the requester in relation to his need are not retrieved because the searcher adheres to the stated request.

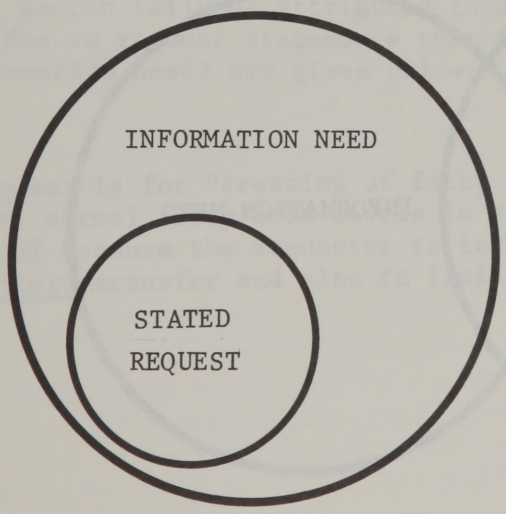


Figure 9

A precision failure due to defective interaction implies that the stated request is more general than the actual information need (Figure 9). Articles of no value to the requester are retrieved. These match his stated request but are of no value because of some additional limitation or requirement that was not given in the request statement.

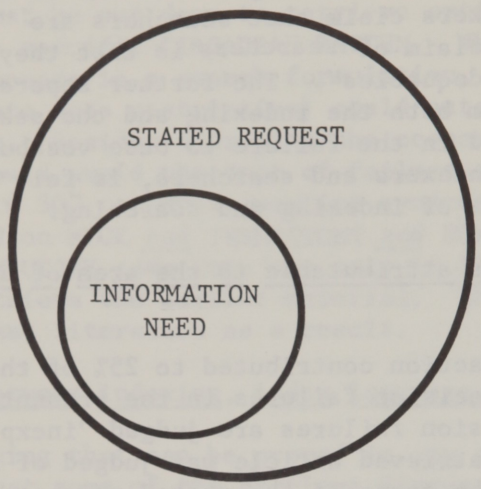


Figure 9

In some searches there is a partial overlap between stated request and information need (Figure 10) and in these cases it is likely that both recall and precision failures will result from inadequate interaction.

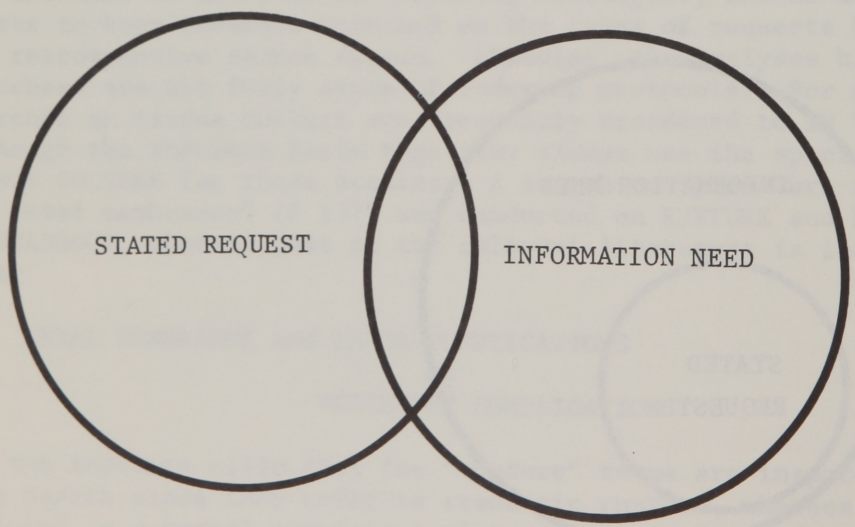


Figure 10

There appear to be basically two types of interaction failure:

1. The situation, long known to librarians, of the user who puts an imperfect request (i.e., a request that does not precisely match his information requirement), or the situation in which the information need is captured imperfectly by a librarian or search analyst.

2. The situation in which the user puts a request that is a fair reflection of his information need, but recall failures result from the fact that he is not fully aware of the types of article that exist and could be of use to him. This type of failure can only be solved by a browsing or iterative search. About 20% of the MEDLARS test searches involving inadequate interaction were judged to be of this type (i.e., the situation in which the requester could never precisely define his need except through some browsing in the literature).

In this evaluation program, a discrepancy between a stated request and an information need has been determined on the basis of:

1. The requester's relevance assessments, particularly the reasons given for judging certain articles of value and others of no value.

2. Revised request statements supplied by the requester in a covering letter to the Evaluator or on the test form Revised Statement of MEDLARS Request.

3. The requester's Record of Known Relevant Documents.

4. In a few cases, by telephone contact between the Evaluator and the requester.

Some examples of search failures attributed to defective interaction (i.e., failures due to request statements that imperfectly represent the area of information need) are given below:

5 The request is for "crossing of fatty acids through the placental barrier; normal fatty acid levels in placenta and fetus". Recall was only 25% because the requester is interested in the broader area of lipid transfer and also in lipid levels in the newborn infant.

11 This search was a complete failure, retrieving 1167 articles of which only one was relevant. The request statement was "cancer in the fetus or newborn infant" but the area of need appears to the relationship between teratogenesis and oncogenesis at the cellular level.

24 The request relates to the metabolism of steroids in liver or liver disease, but the requester did not indicate that animal studies were of no interest.

32 This is an example of a request too specific in one aspect, insufficiently specific in another. The request is for "homonymous hemianopsia in visual aphasia", but the actual area of interest encompasses homonymous hemianopsia and other visual field defects in patients having aphasia as a result of a tumor or cerebrovascular accident.

52 The request reads "skin grafts in monkeys". Only homografts are of interest.

96 The request is for structure and function of the lymphatic system of the lung of any animal. The fact that pathological conditions are not of interest was not made explicit.

115 The request asks for certain specific intestinal microorganisms causing diarrhea or dysentery in cases of protein deficiency or Kwashiorkor. However, the requester seems interested in any studies on the role of infection in protein deficiency.

124 Although the request asks for "changing incidence and mortality from septicemia," only human, adult septicemia is of interest. Many of the retrieved articles deal with fetal infections, the newborn, or animal studies.

147 The request asks for "sodium and potassium ions present in whole blood and erythrocytes", but the requester is interested only in analytical methods and not clinical values.

162 The requester is interested in autoimmune antibody formation against irradiated tissues, and is not interested solely in X-irradiation as stated in the request.

165 The request relates to morphological changes associated with muscular exercise. However, it fails to state that pathological conditions are not of any interest (i.e., the interest relates only to normal muscle), and also to make clear that the requester is interested also in muscle protein biosynthesis and the development of muscle during embryogenesis.

191 The request asks for viruses isolated from human fetal deaths and premature births, but the broader interest is in viruses isolated from children infected in utero.

259 The request relates to brain tissue electrolytes. However, the user wants only in vivo values of brain tissue electrolytes under normal physiological conditions, and conditions of concussion or experimental cerebral edema. He is not interested in other pathological conditions, nor in enzymology, nor in observations on electrolytes in tissue cultures of the brain (incubated brain slices).

260 The request asks for reticuloendotheliosis of the retina or brain. Reticulum cell sarcoma of the eye in general is of interest.

454 The request asks generally for metastases from breast cancer to bone or bone marrow, but the user is really interested only in (1) diagnosis of metastatic breast cancer by bone biopsy, and (2) statistics on incidence of breast cancer metastasizing to bone. He is not interested in therapy.

457 The requester is not interested in all "neurological complications of kidney diseases", but only in CNS changes as a result of uremia.

458 The requester is not interested in all exercise effects on the respiratory and cardiovascular systems (as implied in the request), but only in young, healthy, adult males. He is not concerned with diseased conditions.

473 The request states "neurological and muscular complications of chickenpox and varicella". This should have been stated as neuromuscular complications.

474 Although the requester asks only for atony or ileus, he is interested in any intestinal sequelae of vagotomy or pyloroplasty.

In the design of the test, we hypothesized that, of the modes of interaction presently used within MEDLARS, the most desirable would be the personal interaction mode (i.e., the situation in which the requester visits a MEDLARS center and discusses his needs personally with a search analyst). The other modes (mailed request direct to the system, and the request mailed through a local librarian) would be likely to be less successful on the whole.

To determine the effect of the three modes of interaction, a breakdown of the performance figures for each mode was undertaken. The results are presented in Table 14. These results are completely different from those expected at the time the study was being designed. The performance figures for the 46 "no local interaction" searches (based on requests mailed directly to NLM from the requester) are better than the performance figures for the "local interaction" group, and the performance for the 144 local interaction searches is marginally better than the performance for the personal interaction group.

However, in Table 14 "local interaction" merely means that the request was submitted to NLM by a local librarian acting on behalf of the requester. Although, on the whole, it was very difficult to get librarians to indicate clearly how much, if any, interaction took place locally, the author has split the 144 "local interaction" searches into two groups: (a) a group of 65 searches in which it appears that the librarian merely transmitted a request statement formulated solely by the requester (in some cases the request forms have been completed by the requester with the librarian typing in the library identification later), and (b) a group of 79 searches in which the local librarian appears quite definitely to have interacted, usually by some interview process, with the requester, thereby influencing the request statement as submitted to NLM. The separate results for these two groups are presented in Table 15.

The trend noticed in Table 14 is still quite evident: the performance for those requests in which the librarian acted only as transmitter is clearly better than the performance for those requests in which the librarian participated actively in the request formulation. We can now add the 65 librarian-transmitted requests, involving no local interaction as far as we can tell, to the "no local interaction" group, leaving only 79 searches in the "positive local interaction group". Based on this regrouping, the results are as shown in Table 16. Clearly, the "no interaction" group still outperforms the others. However, on this recalculation, the "personal interaction" group has performed slightly better than the "local interaction" group.

The implications of these results are pregnant. It appears that the best request statements (i.e., those that most clearly reflect the actual area of information need) are those written down by the requester in his own natural-language, narrative terms. When he comes

to a librarian or search analyst, and discusses his need orally, a transformation takes place and, unfortunately, the request statement captured by the librarian or searcher is a less perfect mirror of the information need than the one prepared by the requester himself in his own natural language terms.

Table 14

Breakdown of performance results by mode of interaction

	Precision ratio	Recall ratio
Personal interaction (109 searches)	49.3%	56.4%
Local interaction (144 searches)	49.8%	57.5%
No local interaction	54.8%	61.1%

Table 15

Breakdown of "local interaction" results into "positive" and "negative".

	Precision ratio	Recall ratio
Local interaction: positive (local librarian probably interacted with requester) (79 searches)	46.9%	55.0%
Local interaction: negative (local librarian probably acted only as transmitter, and did not interact with requester) (65 searches)	53.2%	60.6%

Table 16

Recalculation of performance results by mode of interaction.

	Precision ratio	Recall ratio
Personal interaction (109 searches)	49.3%	56.4%
Positive local interaction (local librarian apparently interacted in some way with requester) (79 searches)	46.9%	55.0%
No local interaction (request came by mail direct to MEDLARS or was transmitted by local librarian without interaction) (111 searches)	53.9%	60.8%

Of course, there are many other variables that could be affecting the results in Tables 14, 15 and 16. For example, all the MEDLARS center searches are "personal interaction" while NLM searches are mostly non-personal. The results could indicate that the overall NLM performance is better than that of the other MEDLARS centers. The results could also be influenced by subject field and by type of organization submitting the request. To determine what was influencing what, each of these variables was held constant in turn, while results were tabulated by mode of interaction.

In Table 17 the processing center is held constant. The overall performance figures for the 198 searches processed at NLM are 50.9% precision and 57.9% recall. However the figures are calculated, the "no local interaction" searches came out noticeably better than the "local interaction" group and both outperformed the personal interaction searches (although there are only eight in this case), especially on the recall side.

Table 17

Breakdown of performance figures by mode of interaction for 198 searches conducted at NLM.

	Precision ratios	Recall ratios
OVERALL NLM FIGURES (198 searches)	50.9%	57.9%
PERSONAL INTERACTION (8 searches)	48.3%	45.2%
* LOCAL INTERACTION (request submitted through local librarian: 144 searches)	49.8%	57.5%
* NO LOCAL INTERACTION (request by mail direct to MEDLARS center: 46 searches)	54.8%	61.1%
POSITIVE LOCAL INTERACTION (local librarian apparently interacted in some way with requester: 79 searches)	46.9%	55.0%
* NO LOCAL INTERACTION (request came by mail direct to MEDLARS center or was transmitted by local librarian without interaction: 111 searches)	53.9%	60.8%

* The two groups of bracketed figures are alternative calculations based on the same data.

Table 18

Breakdown of performance figures by mode of interaction for 147 requests submitted by academic organizations.

Mode of interaction	Precision ratio	Recall ratio
PERSONAL (85 searches)	48.2%	57.6%
LOCAL INTERACTION* (44 searches)	47.2%	61.8%
NO LOCAL INTERACTION (18 searches)	55.8%	70.4%

Table 19

Breakdown of performance figures by mode of interaction for 110 requests related to DISEASE.

Mode of interaction	Precision ratio	Recall ratio
PERSONAL (36 searches)	51.6%	62.5%
LOCAL INTERACTION* (54 searches)	44.3%	54.5%
NO LOCAL INTERACTION (20 searches)	52.0%	69.0%

* Unless otherwise stated, "local interaction" means that the request was processed through a local librarian. It does not necessarily imply that the librarian actually influenced the request locally.

In Table 18 the requesting organization is held constant. There were 147 requests from academic organizations (this is the largest group by type of organization). Performance for the "no local interaction" set of 18 searches is greatly superior to the performance of the other types of interaction, but there is a cross-over in the results from the other groups: the personal interaction achieves a higher precision than the local interaction searches, but a lower recall.

In Table 19, the subject field is held constant. The largest subject category was DISEASE with 110 searches. For this group the "no local interaction" searches are clearly superior to the others as far as recall goes; there is very little difference in precision between these and the personal interaction group. For the DISEASE requests, the personal interaction mode has outperformed the local interaction mode.

However we tabulate these results (i.e., whichever variable we hold constant) they always indicate a superior performance for the "no interaction" group of requests. This superiority is more pronounced for recall than it is for precision, which implies a tendency, on the part of search analyst or librarian, to make the scope of the request rather more narrow than it should be. There is not such a clear distinction between the performance figures for the "local interaction" and the "personal interaction". In some tabulations one outperforms the other, while in alternative tabulations the situation is reversed. When we break down the "local interaction" searches into (a) positive interaction, and (b) local handling only, and consider the second group as part of the "no interaction" searches, the gap between "no interaction" and "local interaction" widens.

Another interesting analysis results when we take the group of searches in which interaction failures are known to have occurred, and divide the group up by the mode of interaction in which requests were handled. The results are presented in Table 20. It can be seen from this tabulation that slightly more than 50% of all the searches handled in the personal interaction mode contain recall and/or precision failures attributed to inadequacies in the interaction. This compares with 41.0% for the local interaction group and 43.5% for the "no interaction". However, when we group the "negative local interaction" with the "no interaction" group, it is found that only 37.8% of all these searches contain failures due to user-system interaction, while 46.8% of the true local interaction searches contain failures of this type.

One final analysis presents further evidence to confirm these findings on interaction. Each requester, having seen the complete search printout and also the sample articles, was invited to rephrase more precisely the statement of his request, on a form Revised Statement of MEDLARS Request (see Appendix 2) when he felt, from the search results, that a rewording was necessary (the implication being that the search results had not given him exactly what he expected from the search). If he felt that the original rewording had been adequate, he returned

Table 20

Failures (recall and/or precision) attributed to the area of user-system interaction: breakdown according to mode of interaction in which request handled. Number of searches involved: 134 (44.4% of the total of 302 searches).

	Mode of interaction	Number of searches involving interaction failures	Percentage of total of searches involving interaction failures	Percentage of total of searches handled in this mode
	PERSONAL	55	41.0%	50.5%
*	LOCAL	59	44.0%	41.0%
*	NONE	20	15.0%	43.5%
*	LOCAL - POSITIVE	37	27.6%	46.8%
*	NO LOCAL INTERACTION	42	31.3%	37.8%

* These two pairs are alternative displays of data from the same set of searches.

the form so marked. The results from this analysis will be mentioned again later. For the present, it is sufficient to note that 236 requesters completed the Revised Statement of MEDLARS Request, and 72 of these (approximately 30%) did re-write their request statement. Among the 236 requesters were 82 who submitted their requests by personal visit to a MEDLARS center. It is noteworthy that 34 (41.5%) of these revised their request statements, whereas only 38 (24.7%) of the requesters submitting by mail (through a local librarian or directly to NLM) felt it necessary to rephrase their requests.

While these results are certainly not what we expected when we were designing the test program, they came as no surprise to the author. The search analyses left a very strong impression of interaction failure in the "personal interaction" mode. It appears crucial to the success of a MEDLARS search that the requester be required to write down, in his own natural language, exactly what type of literature he is looking for. When he makes a personal visit to a MEDLARS center, we do not normally have the benefit of this written, natural language statement. Rather, the requester is invited to discuss his need with a search analyst. Unfortunately, at this point, his information need tends to get distorted. The problem appears to be at least partly due to the fact that the requester's need is discussed in terms of, and unduly influenced by, Medical Subject Headings. When the requester is writing down his request, he is forced to think of what exactly he is looking for. In this, he is not particularly influenced by the logical and linguistic constraints of the system. When, however, he approaches a MEDLARS center, if he has not already gone through the discipline of writing down his request, he has a less well-formed idea of what he is seeking (i.e., of the scope and constraints of the search). When this somewhat imprecise need is discussed with a search analyst, in terms of Medical Subject Headings, it tends to become forced into the language and logic of the system. The final "request", rather than representing what the requester wants, represents what he thinks the system can give him, phrased in a way that the system will search for it. In many cases the "request", as recorded by a search analyst, is not a true request at all (at least it resembles nothing that a requester would submit in his own natural language terms). Rather, it is a "pseudo-Boolean statement": a string of MeSH or MeSH-like terms put together in some relationship.

Some examples will illustrate this point:

17 This "request" came out as the single MeSH term CELL DIFFERENTIATION. The requester is interested only in cell differentiation in relation to teratogenesis and carcinogenesis.

27 As recorded by the search analyst, the request came out as: "Wound healing related to radioautography with isotopes of sulfur, estrogens, progestational hormones, mast cells, uterus".

This is not like a request that a doctor would make in a natural language statement. In fact, it is a group of MeSH terms in an implicit logical product relationship (WOUND HEALING and . . .). It is too restrictive in one sense (the requester appears interested in most aspects of the biochemistry of experimental wound healing) and too general in another (only experimental wound healing is of interest). The recall estimate for this search is only 28.6%.

57 "E coli and lipopolysaccharides" is not a request statement but two MeSH terms in a logical product relationship. The requester is not interested in everything on the relationship between the two. He is interested only in lipopolysaccharides in the structure and composition of the cell wall of E coli, and in the biosynthesis of lipopolysaccharides by E coli.

59 "Heart function tests, hemodynamics, capillary and vascular resistance, pulmonary and lung circulation, plasma expanders, hypotension in cases of pneumonia (all types), septicemia, or toxic shock". Again, not a natural-looking request, but a string of terms. This search achieved only 40% recall.

68 "Dr. D is interested in various aspects of the decidua. She is interested (as related to decidua) in: uterine physiology and contraction, labor and premature labor, but not in uterine inertia. She is interested in uterine disease and endometritis, but in none of the other specific uterine diseases, in abortion and abruptio placentae, in pregnancy toxemias or hypertension. She is interested in decidual necrosis, inflammation, capillary permeability, leukotaxin, leukocytes, mast cells, in histamine but not histamine liberation, in chemotaxis, and any effects of decidual homogenates. Also histo- or cytochemistry as related to human studies only but not metabolism or decidual cells, or general cytology or pathology".

Again, a request largely in MeSH terms and, despite its length, not apparently a true reflection of the requester's need. From her relevance assessments and comments, the requester appears interested in pathological conditions (degeneration) of the decidua in human pregnancy, particularly necrosis and inflammation. She is not interested in animal studies (unless there is some direct analogy with a human problem) and definitely not in neoplasms.

78 "Dr. B is interested in articles relating to phagocytosis and the reticuloendothelial system, also in material relating to the RES, leukocytes, phagocytosis and macrophages and bone regeneration, resorption, osteogenesis, bone diseases (as specified), hypo-hyper-pseudo-and pseudopseudohypoparathyroidism, collagen disease (as specified), carbon particle uptake, fluorescent dye uptake, or bacteria (only in experimental studies in animals). The material is limited to English and French. Material relating to the RES, leukocytes, phagocytosis, or macrophages and tobacco, smoking, nicotine, Vincent's infection, and complement has been requested in any language".

This is a classic. It is difficult to imagine what the requester is looking for. It is also difficult to know whether the searcher understood the requester's need or merely got him to agree to a number of MeSH terms. The latter appears more likely: the search retrieved only seven of the fifteen articles named in advance by the requester, and only 47.6% of the 529 articles retrieved were judged of any value.

125 This is a search on enzyme composition of platelets. Apparently the search analyst "persuaded" the requester to limit the enzyme list as given in MeSH. This was a mistake: the requester is interested in all enzyme work on platelets, and recall was only 42.9%.

These examples illustrate some of the problems that tend to occur when requesters are led directly into the language of the system, without first formulating an exact statement of what precisely they would like to get out of the MEDLARS search. A similar situation is likely to occur when a requester confronts his local librarian. Unless the requester is first required to write down a narrative statement of need, this requirement may well become distorted in an oral interview between requester and librarian, and the "request", as recorded by the librarian, thus becomes an imperfect representation of what the requester is seeking. It is for this reason that MEDLARS performs, on the average, less well for searches in which the local librarian has participated actively in request formulation than for searches in which the librarian acts as carrier only. It is not unlikely that the worst local interaction will take place when the librarian uses Medical Subject Headings to help the requester to formulate a request statement.

Moreover, the worst mailed requests are those that the requester phrases in term combinations that he believes the system will search on (i.e., a pseudo-Boolean statement). Such a request (# 185) as: "Body heat and body temperature only as related to perspiration, water vapor, and inert gases. Biological only. Inert gases as related to metabolism, perspiration", is not at all clear, and should not be accepted by the system, because it will inevitably result in an extremely poor search. The search formulation for this request indicates clearly that the searcher did not have a remote idea of what the requester wanted. This is hardly surprising. The requester is really concerned with physiological responses to rapid temperature changes, and with the "thermal comfort" of astronauts in the inert atmospheres of space cabins. If he had stated this in his own terms, and avoided trying to "keyword" his request, we would have had a statement of need that could be transformed into an appropriate searching strategy. As it is, the recall ratio was only 14.3% (1/7) for this request, while only 12.5% of the 250 retrieved citations were of any relevance, and most of the articles retrieved were very far removed from the requester's field of interest (e.g., metabolic studies involving yeast; nitrogen metabolism of diabetic

rabbits; effect of chemical stimulation on the nasal mucosa of rabbits; and sheep wool follicle metabolism).

As far as the personal interaction searches go, there is one factor that has not yet been mentioned, but has to be borne in mind. This is the fact that a search analyst may deliberately record a request more general than the requester would put himself, because the searcher knows that limitations of the vocabulary will not allow a search at the level of specificity required. This could have occurred, for example, in the negotiation of search # 94. The request is recorded as "myringotomy in otitis media" but the requester is interested only in acute purulent otitis and not in serous otitis. However, it is difficult to see how the two could be separated when only OTITIS MEDIA exists in the vocabulary.

If this policy is followed by a searcher, it would seem to be a mistaken one. The request should be recorded at the level of specificity demanded by the requester, and not at some more general level determined to be the level of capability of the system. Otherwise, request and search analysis has no possible value as an indicator of vocabulary inadequacies. Although this kind of generalization could have occurred in a few searches, the analyses have shown that this is not the primary cause of interaction failures at the local level. In fact, most of the additional precision required by a requester, but not included in a request statement (e.g., experimental only, human only), is within the capabilities of the system. In any case, the results show that the local interaction searches as often as not err in the opposite direction. That is, the search analyst records a statement more precise than the actual information need, or a statement that only partly overlaps the need, and recall failures rather than precision failures occur.

The interpretation of these results may be attacked from another viewpoint; namely, that the "no interaction" requests are those that require no interaction and are therefore, in some sense, the "easiest" requests. This argument is invalid. The mode of interaction in this evaluation has no direct relationship to complexity of requests. Personal interaction will occur when it is convenient for a requester to visit a MEDLARS center. If he is not situated in or close to a MEDLARS center, his request will usually arrive by mail. In some cases the request will come directly to NLM; in others, it will be transmitted through a local librarian. In this evaluation program, there appears to be no direct relationship between complexity of requests and the action of the local librarian.

Some of the participating librarians, as a matter of policy, interview the requester orally and "interact" with him. Others, as a matter of policy, merely record verbatim the requester's statement, or merely transmit his pre-written statement. The results show that the "interaction" requests, on the average, are worse than those merely transmitted by the librarian.

In passing, it should be noted that some requesters themselves have been made aware of the dangers inherent in forcing their need into the language of the system. The following comments, included in covering letters to the author or on the Revised Statement of MEDLARS Request, are highly pertinent:

1. "The actual topic of my project is prognostic factors in the surgical treatment of temporal lobe epilepsy. Because of a lack of fine headings or subtitles that could be used in the formulation of the search request, the request was finally formulated as factor analysis of temporal lobe epilepsy."

2. "A large number of the references supplied by MEDLARS were irrelevant. However, with the titles or subject headings we had to choose from, I could not be more specific. I think that a descriptive paragraph by the requester might be more effective than letting him draw up the headings from the list available. Then MEDLARS would know more precisely what the investigator wanted."

3. "The difficulty remains in having a limited number of categories to choose from. I believe it might be better if the individual requesting a search could write a descriptive paragraph of what he wants and let your experts choose the categories".

Improving request statements.

The large number of failures attributed to the area of user-system interaction, prompted the conduct of a detailed analysis of the searches involved. This analysis was intended to determine what data, if it could have been collected from the requester, would have helped to define more exactly the area of his actual information requirement. Put differently, can we design a search request form that, without being too complex, will assist the requester in defining the exact scope of his need, and at the same time will give the searcher data adequate for a proper understanding of the requester's need?

The following observations resulted from this analysis:

1. Obviously, the prime requirement is a complete statement of what the requester is looking for. This should be a statement in the requester's own natural-language, narrative form. It must not be deliberately phrased in the language of the system. It must not be artificially structured in a form that the requester believes will approximate to a MEDLARS search strategy.

2. Knowing the purpose of a search would clearly have helped, in a number of cases, to precisely define its scope. We only know of these cases because the requester has told us the reason for his search after it was completed. It is not unlikely that this data would have helped to define the scope of many of the other searches. For

example, the request for search # 11 came out as "cancer in the fetus or newborn infant", while the actual area of interest covers "the relationship between teratogenesis and oncogenesis at the cellular level". The search might have been successful had the requester informed us, as he did post facto, that he had been asked to write an article, for Advances in Teratology, on "some aspect of cell differentiation or malformations and how they are related to the cancer problem." Similarly, search # 495 relates to the effect of inactivity (bed rest or cage rest) on the circulatory system or on calcium or bone metabolism. Recall could have been improved had the searcher known something of the requester's research project, which relates to parallels between the cardiovascular effects of bed rest or cage rest and the "postural intolerance" caused by confinement in space capsules. Knowing this, the searcher might well have incorporated SEALED CABIN ECOLOGY and/or POSTURE into the search formulation, thereby retrieving a number of missed articles directly related to "cardiovascular conditioning" in space cabin confinement.

Likewise, in search # 492, the requester expressed an interest in sulfur, resorcinol, hexachlorophene and allantoin as topical medicinal agents in various combinations. However, he failed to indicate that his interest lies in the use of these as topical medicinal agents in the treatment of acne. The searcher could not, therefore, be expected to include ACNE and DRUG THERAPY in the formulation. Consequently, both recall and precision suffered.

Knowing the purpose of a search may help to establish the recall and precision requirements and tolerances of users. A physician writing a book on "connective tissue in ocular disease", and asking for a search on this subject will demand high recall and tolerate low precision. That is, he will be prepared to examine a comparatively large number of nonrelevant citations to assure himself that he is not missing any articles of prime importance. On the other hand, we have a requester who is preparing a seminar, for nurses, on kidney transplantation. He does not necessarily want everything, but would like to see a number of recent major surveys of the entire subject. The requirement is for high precision, and a low recall will be tolerable.

3. The titles of articles that the requester knows to be relevant at the time he makes his request (collected in the present evaluation program on the form Record of Known Relevant Documents) can be of great potential value in helping to define the scope of a search. For example, search # 5 relates to "crossing of fatty acids through the placental barrier; normal fatty acid levels in placenta and fetus". The fact that the requester is really concerned with a broader topic (crossing of lipids through the placental barrier; normal lipid levels of placenta, fetus, or newborn infant) is well illustrated by the titles supplied in advance by him:

> "Release of free fatty acids from adipose tissue obtained from newborn infants."

"Changes in the nucleic acid and phospholipid levels of the liver in the course of fetal and postnatal development."

"The amount of lipids and proteins in the foetus of rats and rabbits and in newborn infants."

In search # 33 the request was for "pulmonary nocardiosis", while the requester appears interested in the whole field of nocardia infections, as the following titles seem to indicate:

"Human nocardiosis: a clinical picture."

"Rapid differentiation between nocardia and streptomyces by paper chromatography."

Recall was low in search # 165 because the requester asked for "morphological changes associated with muscular exercise" and did not make it clear that he was also interested in muscle protein biosynthesis and in the development of muscle during embryogenesis. His Record of Known Relevant Documents, on the other hand, contains the following titles:

"Ultrastructure of developing muscle cells in the chick embryo."

"Amino acid incorporation into protein by cell-free preparations from rat skeletal muscle."

The classic request for "inert gases in relation to perspiration" (# 185) might well have resulted in a much better search if supplemented by the following titles supplied in advance by the requester:

"Gas transfer across the skin of man."

"Thermal comfort zones for helium-oxygen atmospheres."

In search # 191, the request was for "viruses isolated from human fetal deaths and premature births." A better statement would be "viruses isolated from children infected in utero", as the following titles indicate:

"Rubella virus carrier cultures derived from congenitally infected infants."

"Viruses in cell cultures of kidneys of children with congenital heart malformations."

Note that titles of "known relevant" articles will be most useful in indicating a request more specific than the actual information need. If the titles quoted are obviously outside the scope of the stated request, this is an indication of a request statement that is too precise.

On the other hand, if the titles supplied are more specific than the stated request, this does not really tell us very much - since they are included within the scope of the search as defined by the request statement. Sometimes, however, the cited titles will relate to a topic much more precise than the general area covered by the request statement, and this will tend to identify a request that is probably too general in relation to the precise information need. For example, the requester of search # 11, on "cancer in fetus or newborn infant", could name only one known relevant paper before the search, namely: "Association of Wilm's tumor with aniridia, hemihypertrophy and other congenital malformations". The fact that the requester could name only one relevant article, and that so much more specific than the statement of his need, surely suggests that the request statement is much broader than it should be.

We are not, of course, advocating that search formulations be based on titles of known relevant papers rather than request statements. The two complement each other. The "known relevant" titles are most useful in signalling the fact that the request statement is probably imperfect and needs further clarification.

4. Occasionally, the requester's estimate of the volume of the relevant literature published on the subject of his request, within the time span of MEDLARS, will also signal a request statement that may either be too broad or too narrow. The danger here is that the requester may base his estimate on the request statement rather than on the volume of literature he expects on the precise topic of interest.

5. One of the problems is to get the requester to be sufficiently precise when he makes a narrative statement. We do not want him to omit any aspect of interest, but we would like to be able to eliminate certain categories of articles that will be of no possible use to the requester. In other words, we want to know what we can definitely exclude from a search. It is difficult usually for the requester to think in terms of exclusions when he prepares his narrative statement of need. Therefore we must help him to usefully delimit the scope of the search.

This could be done, for example, by incorporating into the request form a brief, carefully designed questionnaire relating to the previously recorded narrative statement. Certain types of exclusions, or limitations of scope, are applicable to many MEDLARS searches. For example, some requesters are interested only in humans, others are interested in both human and animal studies. Some requesters are interested only in particular sexes or age groups (e.g., "young, healthy, adult males"). Experimental studies are the sole concern of some requesters, while clinical case histories are all that is wanted by others. Some will accept all clinical studies, including single case reports, while others are only interested in "large series" of cases. Again, one requester will be interested in a particular organ under normal physiological conditions, while another will be concerned with pathological conditions of the

same organ, or in certain pathological conditions only (e.g., not neoplasms).

Such generally-applicable limitations on a search could easily be incorporated into the demand search request form, either in the form of direct questions or by the use of check-off boxes (these would resemble the "check-tags" appearing on the MEDLARS indexing Data Form for generally applicable descriptors; e.g., age groups, sex, experimental animals, type of study). Very many of the request statements could have been made much more exact if the requester had been given this assistance in defining the scope of the search. For example, in # 5 newborn infants are of interest as well as fetuses; in # 8, clinical case studies are of no interest whatsoever (only 5% precision in 300 items retrieved); in # 24, animal experiments are of no interest; in # 27, the requester wants only experimental wound healing; in # 32, the requester is interested only in aphasia resulting from a tumor or cerebrovascular accident; in the search relating to the decidua (# 68), animal studies are of no interest, and neoplasms are also unwanted; in # 124, the requester is not interested in all septicemia, but only in human, adult septicemia; in # 126, the requester is interested in erythrocytic hematological findings, physiological or pathological, in human babies; in # 147, analytical methods relating to sodium and potassium ions are wanted; clinical values are not wanted; in # 165, the effect of exercise on normal muscle is wanted; pathological conditions are of no interest.

Having gone through this questionnaire approach, we should end up with a very precise statement of the requester's information need, including limitations that he would not think to include in his original request statement. We are now able, in our searching strategy, to eliminate a large segment of the data base (within the limitations of the vocabulary) that is of no interest to the requester, allowing us to undertake a broader search in the pertinent segment, while still obtaining a tolerable precision ratio.

6. The request form should also be designed to determine the recall and precision requirements and tolerances of the requester. At its simplest, the form could merely ascertain whether the requester would like all papers making some reference to the subject matter of interest, or whether he wants only papers in which the subject matter is treated centrally. Alternatively, it could be made more complex by allowing the requester to choose from a number of alternatives. For example:

"Which would you prefer:

1. A search retrieving about 60% of the relevant articles within MEDLARS, but with about 50% irrelevancy in the search?

or

2. A search retrieving about 90% of the relevant articles within MEDLARS, but with about 80% irrelevancy in the search?"

Figure 11

General outline of a proposed form to record request for literature search.

1. Requester: name, title, address, telephone number
2. Statement of need in narrative form
3. Statement of purpose of search
4. List of up to five known relevant documents, preferably published within the time span of the system.
5. Brief questionnaire designed to usefully delimit the scope of the search. For example:

   Is the requester interested in the subject only in relation to humans, is he interested only in veterinary medicine, or is he interested in both?

   Is he interested in normal physiology or in pathological conditions?

   Is he interested in animal experiments, in clinical research, in *in vitro* studies?

   Would case reports be of interest -- either large clinical series or individual case studies?

   Are there any age limitations, sex limitations?

   Are there any language restrictions?

   Are there any other restrictions; for example, by race, by geography, etc.

6. What are the recall and precision requirements or tolerances of the requester?

REQUEST

Heart preservation methods.

SEARCH FORMULATION

(2)

[Heart] AND [Preservation, biological
Tissue banks
Transplantation
Transplantation, autologous
Transplantation, homologous
Transplantation, heterologous
PERfusion
REFRIGERATION
Storage]

OR

[HEART TRANSPLANTATION] of only limited value

TIME EXPENDED

Please record the amount of your time spent in examining and evaluating this proposed search strategy:

5 minutes

Figure 12

REQUEST

Effects of pulmonary infection on the kidney and on the development of renal disease.

SEARCH FORMULATION

(1)

Respiratory tract infections or
~~Lung diseases or~~
Lung abscess or
Lung diseases, fungal or
Lung diseases, parasitic or
Echinococcosis, pulmonary or
Pneumonia or
Bronchopneumonia or
Pleuropneumonia or
~~Pneumonia, lipid or~~
Pneumonia, lobar or
Pneumonia, rickettsial or
Pneumonia, viral or
Tuberculosis, pulmonary or
~~Pneumonia~~, interstitial plasma cell

AND

(2)

Kidney diseases or
Acute renal failure or
Anuria or
~~Diabetic nephropathies~~ or
~~Kimmelstiel-Wilson syndrome~~ or
~~Hydronephrosis~~ or
~~Kidney calculi~~ or
Nephritis or
Nephritis, interstitial or
~~Nephrocalcinosis~~ or
~~Nephrosclerosis~~ or
Nephrosis or
Nephrosis, lipoid or
Glomerulonephritis or
~~Goodpasture's syndrome~~ or
Nephrotic syndrome or
~~Perinephritis~~ or
~~Pyelitis~~ or
~~Pyelocystitis~~ or
~~Pyelonephritis~~ or
~~Pyelonephritis, acute~~ or ~~necrotizing~~
~~Rickets, renal~~ or
Uremia or
Tuberculosis, renal or
Hypertension, renal or
Kidney or
~~Juxtaglomerular apparatus~~ or
Kidney glomerulus or
~~Kidney pelvis~~ or
~~Kidney tubules~~ or

BUT NOT

(3)

any term indicating an animal study

TIME EXPENDED:

Please record the amount of your time spent in examining and evaluating this proposed search strategy:

30 mins

Figure 13

A rough outline of the type of search request form envisaged in the above discussion is given in Figure 11.

Experimentation with modes of interaction

Within the evaluation program, we attempted a small-scale simulation of two modes of interaction (requester validation of proposed searching strategy, and iteration) not presently used in MEDLARS. This experimentation was difficult to undertake because it had to be done without perturbing the operations of the existing system, which was being evaluated in its present form. Of necessity, then, this experimentation had to be very limited in scope.

In the case of searches selected at random for "validation" (but only from the "nonpersonal interaction" searches), the requester was presented with a graphical display of the search strategy before the search was conducted. He was allowed to review this, suggesting possible additions or deletions. To avoid perturbation of the present system, no use was made of the "validated formulation" in revising the search strategy. In fact, it was not seen by the Search Section at NLM. However, analysis, after the requester's assessments had been made, allowed general observations on the probable effect of the suggested changes on the search performance.

A total of 45 proposed strategies were submitted for validation. Replies were received from 30 of these requesters, of which eight indicated satisfaction with the proposed strategy. Of the remaining 22, 12 made certain suggested changes of a somewhat minor nature (i.e., additions or deletions that would have some slight effect on the search but would not result in appreciably different results). Such a "validation" is shown in Figure 12. Here the requester has suggested a few additional "preservation" terms. He has also deleted the single term HEART TRANSPLANTATION, but this is inconsistent because the combination HEART and TRANSPLANTATION is retained.

Ten of the requesters, however, suggested major changes to the proposed strategy, thereby indicating that the original request statement was a very imprecise reflection of what was really wanted. This can be seen, for example, in Figure 13, which indicates quite clearly that the requester is not interested in all effects of pulmonary infection on the kidney.

General observations on this validation procedure are as follows:

1. The validation procedure is usually less valuable for the suggested changes themselves (a requester, being unfamiliar with indexing conventions, will sometimes delete an essential term) than for the significance of these changes in defining the scope of the request. In other words, the formulation as revised by the requester is unlikely to be something we can accept intact. However, it will frequently serve to indicate a gap between stated request and actual information need.

2. Such a technic is cumbersome and time-consuming to apply to a system in which much of the negotiation is by mail or telephone. It might, however, be more applicable to the type of on-line system in which the requester is browsing through the file with the assistance of a trained searcher.

To obtain some idea of the potential benefits of iterative searching within MEDLARS, requesters were invited to re-phrase their requests after examining the set of articles submitted for assessment. An example of a completed Revised Statement of MEDLARS Request is included in Appendix 2. These revised statements were not, in fact, used to generate a revised search. They were merely used for analysis purposes: they helped to establish the distance existing between original stated request and actual information need, and indicated the ability of requesters to restate their needs more clearly after seeing a sample search output.

Although forms were sent to all 302 requesters who completed relevance assessments, the form was returned by only 236 of these. Approximately 70% of the respondents did not feel that it was necessary to revise their original statement, but 72 requesters (about 30% of those completing the form) did in fact produce a revised statement. Of these, only 9 (12.5%) were more general than the original request statement; 61 were either more specific or re-emphasised the specificity of their original request; two revised statements were partly more general than the original, and partly more specific.

These results illustrate the general capabilities and limitations of iterative searching in a system like MEDLARS. On seeing a sample of retrieved articles (or, possibly, a sample citation printout) the requester is prompted to be more precise. In particular, he can exclude certain categories of materials that he did not think to exclude originally (I do not want clinical studies, I do not want animal experiments, I do not want public health aspects). However, after seeing a sample output he cannot be expected to generalize his request, because he really has no basis for doing so. In fact, even after seeing the complete search output (as was the case in our "simulation" of iterative searching), few requesters were able to make their original request more general. The reason, of course, is obvious. The typical requester is aware that certain of the retrieved articles are of no value, and indicates that he would like these excluded. However, unless he is absolutely au fait with the literature, he does not know what the search has missed. Even if he is fairly familiar with the literature, he will not necessarily notice non-retrieved items. If he does notice them, he may not be too upset because they are articles he was aware of anyway.

In other words, while a requester is the only person who can really judge the precision of a search, he is usually in no position to be able to judge its recall because he is not aware of what the system

has missed. Whereas the author has a fairly good idea of how MEDLARS performed in most of the 302 test searches, few of the requesters themselves know how well or how badly the system behaved. In fact, we know that in some searches, with which the requester indicated satisfaction, MEDLARS retrieved only a small fragment of the relevant literature in its base.

The relevant point here is that, while iterative searching could improve the precision of a MEDLARS search, it could only improve recall if the original strategy (i.e., the search formulation producing a sample output that is presented to a requester for his examination) were deliberately made much broader than the scope of the request statement, thus allowing a certain amount of "browsing" by the requester. It is noteworthy that 41.5% of the "personal interaction" requesters revised their request statements after seeing the search results, while only 24.7% of the "nonpersonal interaction" requesters revised theirs. The implications of this have already been discussed.

Recall and precision failures attributable to computer processing

There were ten searches in which recall or precision failures were attributed to the general area of computer processing. Such failures accounted for 1.4% of all the recall failures and 0.1% of all precision failures. A recall failure is attributed to computer processing when the index terms assigned to the missed article match the search formulation used to conduct the search. However, the article does not appear in the computer-printed demand search bibliography. Likewise, a precision failure of this type is one in which the unwanted article does not match the search formulation, yet it is printed in the demand search bibliography. In other words, in both recall and precision failures, there is no defect that can be attributed to the "intellectual" aspects of the system. The problem, therefore, must be due to some aspect of machine processing.

Recall failures of this nature occurred, for example, in search # 16. The recall ratio was 11/16, and the indexing of all five missed articles precisely matched the search formulation (BACILLUS SUBTILIS and BACTERIOPHAGE). When the search was re-run, several weeks later, the previously missed citations appeared in the printout. The causes of these errors have not been precisely identified. They could be due to a slight defect in the programs, to tape problems occurring while the searches are being run, or possibly to file maintenance procedures affecting one of the terms involved in these searches. Because they constitute an insignificant proportion of all recall and precision failures, no special effort was made to determine exact causes.

Precision failures of the "value judgement" type

Seventy-one of the precision failures (2.3%) were due solely to the fact that the requester considered the retrieved article relevant in some way to his request, but insignificant or trivial. Some of these articles were discarded because they contributed very slightly (e.g., a letter to the editor) to a research topic upon which extensive literature exists. Most, however, were discarded on the grounds of "level of treatment". For example, a scientist doing advanced research receives some articles, in the area of his work, that are written for the general practitioner. Similarly, a specialist on neuroradiology receives articles on brain scanning written for the nurse or the technician.

"Inevitable" retrievals

Only four (0.1%) of the precision failures were regarded as "inevitable". These are cases in which the retrieved article is correctly indexed and correctly matches the search formulation. Yet it is "irrelevant" to the request and there is nothing that one could reasonably have done to avoid it.

The novelty ratio

An interesting ratio, from the point of view of search analysis, is the novelty ratio, which indicates what proportion of the articles judged of value by a requester were brought to his attention for the first time by the MEDLARS search. In search # 1, as an example, there were six "major value" articles in the sample assessed. However, these were all known to the requester prior to the MEDLARS search (i.e., the novelty ratio is 0/6). Of the 13 "minor value" articles in the sample, the requester was previously aware of ten (i.e., the novelty ratio is 3/13). The overall novelty ratio of the search is 3/19.

When we consider the novelty ratio in relation to the other data accumulated, it is possible 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. In search # 1, for example, the requester was thoroughly familiar with the literature of his field. He was able to name 17 items before the search (probably only a fraction of the items known to him) and none of the "major value" articles in the random sample were new to him. Likewise, only three of the "minor value" articles were novelty items. It appears that, in this case, the MEDLARS search is conducted to assure the requester that he is not missing articles of central importance, and to bring to his attention, for the first time, certain articles of peripheral relevance to his research topic.

In search #40, although the volume of the literature is much less, the requester was previously aware of the few key items. The function of the MEDLARS search is again to supply the unknown items of peripheral relevance.

On the other hand, the originator of search # 32 was unfamiliar with the literature of his research topic and was probably approaching this particular area for the first time. He was able to name only one pertinent citation ahead of search, and all eight major and 13 minor value articles contained in the random sample were new to him.

FACTORS AFFECTING THE PERFORMANCE OF A MEDLARS SEARCH

Figure 14 is a scatter diagram of the results of 299 test searches for which we have recall and precision figures. It shows no discernible pattern and emphasises the reservations that are necessary when we speak of an average performance of 58% recall and 50% precision. There is, in fact, no single search of that performance level. Furthermore, if we take the area bounded by the average ratios $\pm$ 5% (i.e., 53% to 63% recall and 45% to 55% precision) we find that only four search results fall within these ranges, this being only one more than would have been expected if the results had been completely random.

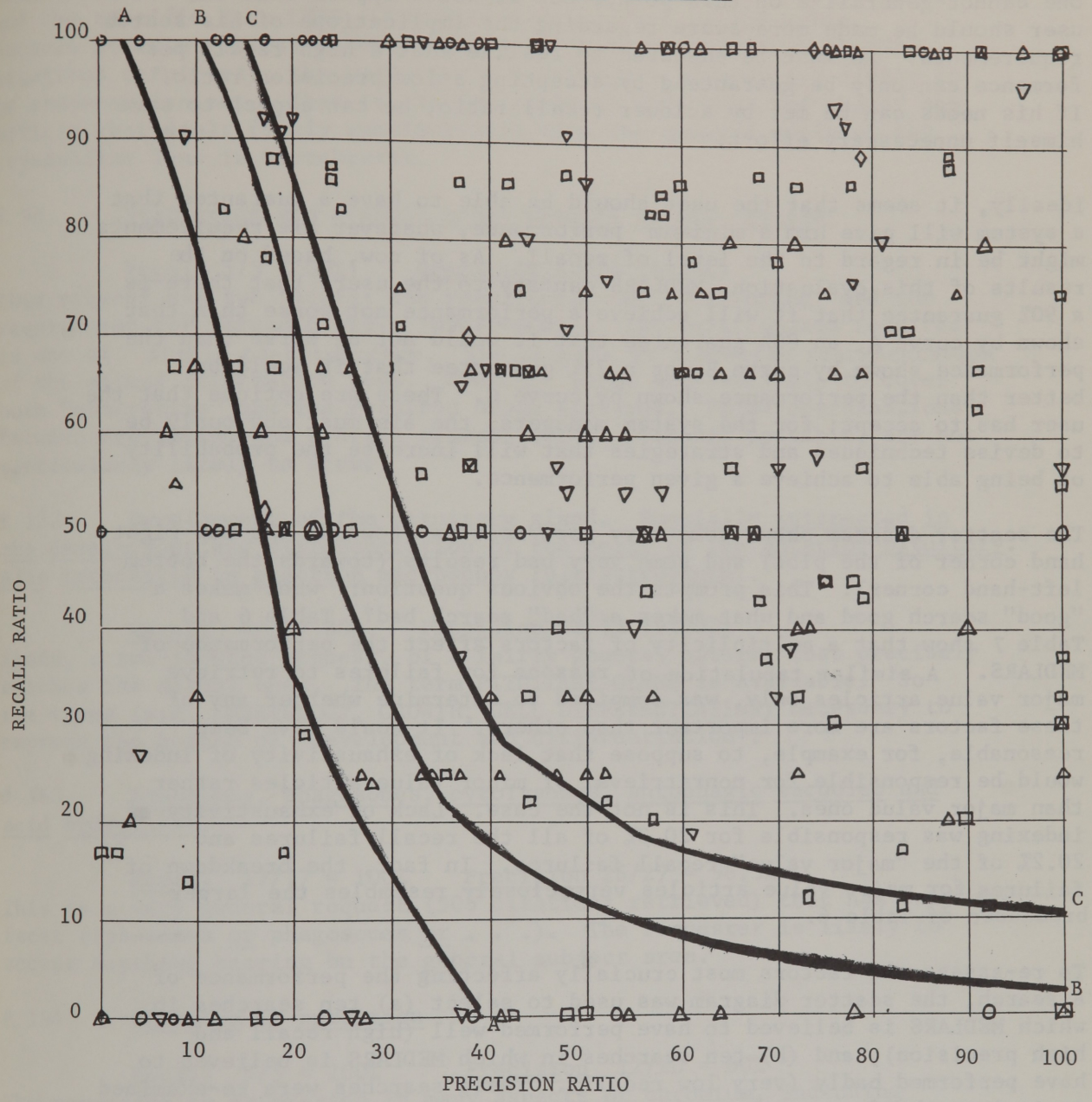
The plot makes clear a number of aspects of the results. It can, for instance, be seen that rather more than two searches out of five result in a minimum recall of 66%; on the other hand, in one search out of five the requester receives not more than one third of the relevant documents. More particularly, however, the plot raises the question - as do all the test results - as to why there should be such a wide variation in search performance. The analysis of failures points to the reasons why non-relevant documents are retrieved and relevant documents are left in the system, but it leaves unsolved the problem as to what can be done to ensure a satisfactory performance for any individual search.

As Calvin Mooers pointed out in a meeting of the MEDLARS Evaluation Advisory Committee, whatever changes might be made in the future, there are now some half-million citations in MEDLARS and it would be some years before a change in, for instance, present indexing policy, could be expected to have any major effect on the overall performance.

The inevitability of the inverse relationship between recall and precision has been known to research workers for some years; what have not been generally appreciated are the practical, economic consequences of this for operating systems. It can be accepted that, in some way or other, it must always be possible to obtain 100% recall. It may not be practical, in the sense that the total number of documents retrieved might be too great, but it always can be done. But is 100% recall always necessary? The typical user of a system such as MEDLARS may be conditioned to the idea that he must have every relevant

Figure 14

Scatter diagram of results of MEDLARS test searches.



No. of documents in recall base	
o	1 - 2
△	3 - 5
□	6 - 10
▽	11 - 20
◇	21 ÷

paper; he is almost certainly not aware of how much irrelevant material he will be required to scan in order to achieve this. Quite clearly one cannot generalize on this point, but it would appear that the user should be made more aware regarding the implications of his search requirements. He must be educated to realize that a high recall performance can only be guaranteed by accepting a low precision ratio. If his needs can be met by a lower recall ratio, he can expect to save himself unnecessary effort.

Ideally, it seems that the user should be able to have a guarantee that a system will give him a minimum performance, whatever his requirements might be in regard to the level of recall. As of now, based on the results of this evaluation, MEDLARS can say to the users that there is a 90% guarantee that it will achieve a performance not worse than that shown by curve A, an 80% guarantee that it would not be worse than the performance shown by curve B, or a 75% guarantee that it would be better than the performance shown by curve C. These are options that the user has to accept; for the system managers, the aim must obviously be to devise techniques and strategies that will increase the probability of being able to achieve a given performance.

The scatter diagram shows some very good results (towards the top right-hand corner of the plot) and some very bad results (towards the bottom left-hand corner). This prompts the obvious question: what makes a "good" search good and what makes a "bad" search bad? Table 6 and Table 7 show that a multiplicity of factors affect the performance of MEDLARS. A similar tabulation of reasons for failures to retrieve major value articles only, was compiled to determine whether any of these factors are more important than others. It would have been reasonable, for example, to suppose that lack of exhaustivity of indexing would be responsible for nonretrieval of minor value articles rather than major value ones. This is not the case. Lack of exhaustivity of indexing was responsible for 20.3% of all the recall failures and 20.2% of the "major value" recall failures. In fact, the breakdown of failures for major value articles very closely resembles the larger breakdown of Table 6.

To re-assess the factors most crucially affecting the performance of a search, the scatter diagram was used to select (a) ten searches in which MEDLARS is believed to have performed well (high recall and high precision), and (b) ten searches in which MEDLARS is believed to have performed badly (very low recall). These searches were re-examined to isolate the factors that most vitally affected the performance level of each. The results are presented below.

Ten "good" MEDLARS searches

1 The crystalline lens in vertebrates

Recall 15/17 (88.2%) Precision 19/24 (79.2%)

This is a request involving only two facets in a simple whole-part relationship. An explosion on VERTEBRATES takes care of one of these, and the specific term LENS, CRYSTALLINE takes care of the other. Because it is a general request with a large body of literature (344 citations retrieved), the requester's relevance standards are likely to be somewhat lenient. That is, he is likely to accept as "of value" any article that deals fairly substantially with the general topic of the crystalline lens in vertebrates.

90 Stuttering and organic brain changes or EEG changes

Recall 5/5 (100%) Precision 21/24 (87.5%)

This request has two facets: the requirement for "stuttering" and the requirement for "organic brain involvement". The relationship required is one of simple co-occurrence. The request is a precise representation of the actual information need. Specific terms are available to cover both "stuttering" and the required brain diseases. False coordinations between STUTTERING and a brain disease term, although possible, are not particularly likely to occur.

137 Development of the pituitary gland. Specially interested in the development and differentiation of the cells of the anterior pituitary. Also interested in electron microscopy (ultrastructure).

Recall 4/5 = 80% Precision 12/13 (92.3%)

Again, a two-faceted request, of a fairly general nature, that precisely matches the actual need. The term PITUITARY GLAND is available. So are terms (e.g, EMBRYOLOGY, GROWTH, CELL DIFFERENTIATION) necessary to express the notion of "development and differentiation".

163 Lysosomes, phagosomes, cytolysomes, autophagic vacuoles and acid hydrolases.

Recall 17/17 = 100% Precision 26/27 = 96.3%

This is a very general request (506 citations retrieved) that has only one facet (lysosomes or phagosomes or . . .). The requester is likely to accept anything bearing on the general subject area.

194 Chromium in nutrition

Recall 5/5 = 100% Precision 12/20 = 60%

The requester is interested in many aspects of chromium, including toxicology, normal levels in tissue, methods of analysis, effect on glucose metabolism and association with diabetes. Again, a fairly broad request. A single term search on CHROMIUM or CHROMATES would have achieved virtually 100% recall with an acceptable 25-30% precision (total of about 200 citations retrieved)

208 Course, treatment and pathology of syringomyelia

Recall 7/7 = 100% Precision 27/32 = 84.4%

This is really a comprehensive search on syringomyelia. It was conducted on the single term SYRINGOMYELIA.

209 Methods that have been used to prepare radio-labeled fibrinogen

Recall 7/7 = 100% Precision 21/24 = 87.5%

This is quite a specific request involving two notions in a precise relationship: a radioisotope used to tag fibrinogen. It was successful because the specific terms FIBRIN and FIBRINOGEN, as well as the radioisotope terms, are available in MeSH. Moreover, there is comparatively little likelihood of a false coordination between FIBRINOGEN and an isotope term (i.e., if both have been used in indexing, it is very likely that the isotope has been used to tag the fibrinogen).

239 Experimental renal hypertension in any animal

Recall 8/8 = 100% Precision 10/15 = 66.7%

This is not a complex request. The term HYPERTENSION, RENAL exists, and merely requires coordination with terms indicating animal or experimental studies.

277 The nature of the renin-renin substrate reaction

Recall 16/16 = 100% Precision 21/23 = 91.3%

This subject is complex and has many ramifications. However, the requester wants a comprehensive search in an area in which much literature exists. Again, his relevance standards are not very stringent: he is willing to accept any article (he expected 200-500 and 500 were retrieved) bearing in some way on this broad subject. The search strategy turned out to be comprehensive, and suitable terms in the vocabulary (RENIN, ANGIOTENSIN, ANGIOTENSINASE) made possible an excellent result.

281 Connective tissue in ocular disease

Recall 7/9 = 77.8% Precision 14/20 = 70%

Another request covering a broad subject area. The requester is writing a book on the subject and wants comprehensive retrieval. He expected over 500 citations, and 343 were actually retrieved. Once more, his relevance standards are not too stringent. This allowed the searcher to formulate a fairly broad search to achieve a high recall.

Ten "bad" MEDLARS searches

123 Life islands in relation to humans

Recall 2/9 = 22.2% Precision 12/23 = 52.2%

This is not a complex request, and it should be well within the capabilities of the system. Recall was low because the specific LIFE ISLANDS came into use as a provisional heading only on 9/25/66. At that time the instruction index under GERM-FREE LIFE appeared in the Authority File. The searcher used both LIFE ISLANDS and GERM-FREE LIFE (on the assumption that the latter term was used for material input

prior to 9/25/66). As it happens, the earlier articles on life islands are indexed under INTENSIVE CARE UNITS and/or ANTISEPSIS. This is a case in which lack of a specific MeSH term, and lack of an entry vocabulary term, has caused indexing inconsistencies and led to searching failures.

128 Growth hormone in newborn infants

Recall 1/6 = 16.7% Precision 10/22 = 83.3%
This should be an easy search. Unfortunately, the request does not fully represent the area of interest. The requester is concerned also with maternal-fetal exchange of growth hormone. Since MATERNAL-FETAL EXCHANGE was not used in the search formulation, recall was inevitably low.

160 Treatment of childhood nephrogenic diabetes insipidus by means of chlorothiazide, hydrochlorothiazide, low sodium diets and adactone.

Recall 1/9 = 11.1% Precision 5/6 = 83.3%
An unsatisfactory result due partly to a request statement more specific than it should be. The requester, from his relevance assessments, appears to be interested in the use of diuretic agents (not just the two mentioned) in the treatment of nephrogenic diabetes insipidus. The requirement for "childhood" is non-essential. He also appears interested in articles on non-nephrogenic diabetes insipidus when the mechanism of action of a diuretic agent is discussed in detail. The searcher was not even successful in retrieving articles that match the request precisely, because of inadequacies in the index language. There is no specific term for "nephrogenic diabetes insipidus" and nothing in the entry vocabulary to say how this is to be indexed. The searcher coordinated DIABETES INSIPIDUS with kidney or kidney disease terms, but indexers appear to use DIABETES INSIPIDUS alone to express the nephrogenic variety of this disease.

162 Auto-immune antibodies produced against tissues damaged or altered by x-ray irradiation.

Recall 2/10 = 20% Precision 7/12 = 58.3%
Inadequate recall because: (a) the requester is interested in autoimmune antibody formation against irradiated tissues (not just x-ray irradiation), and (b) the searcher did not cover all possible approaches to retrieval. AUTOANTIBODIES, ANTINUCLEAR FACTORS and AUTOIMMUNE DISEASES were used, but many additional relevant articles could have been retrieved on other immunology terms (ANTIBODY FORMATION, GAMMA GLOBULIN, ANTIGEN-ANTIBODY REACTIONS) coordinated with radiation terms.

174 Testicular biopsy in infertility and endocrine disease. Also, the effect of surgical and hormonal therapy on sperm count, testicular morphology, and fertility.

Recall 3/11 = 27/3% Precision 10/20 = 50%

This recall failure is partly due to the fact that the searcher made no attempt to cover one complete aspect of the request (effect of surgical and hormonal therapy). Also, the MEDLARS indexing is not sufficiently exhaustive to obtain high recall for this request. For example, consider the relevant article by Valencia overleaf. This was indexed under ABNORMALITIES, CHROMOSOMES*, EPITHELIUM*, SPERMOTOZOA*, and STERILITY. However, the fact that a testicular biopsy was involved was not brought out in the indexing. The searcher could not reasonably have produced a strategy to retrieve this article. Some 50% of the recall failures were similar situations in which the "testicular biopsy" aspect of an article was not covered by the indexing.

177 Obstetric management of the situation of premature rupture of the fetal membranes.

Recall 1/5 = 20% Precision 3/23 = 13%

There is no specific MeSH term for the topic of "premature rupture of the fetal membranes", and nothing in the entry vocabulary to indicate how this is to be indexed. Consequently, there has been no consistency in the indexing of this topic. Some articles are indexed under FETAL MEMBRANES and LABOR, PREMATURE, some under FETAL MEMBRANES and LABOR COMPLICATIONS, some under FETAL MEMBRANES and PREGNANCY COMPLICATIONS, some under LABOR PREMATURE and PREGNANCY COMPLICATIONS, some under FETAL MEMBRANES and RUPTURE SPONTANEOUS. Moreover, where no specific term exists in MeSH, and the notion does not appear in the entry vocabulary, there is a strong tendency for this topic to be omitted by the indexer. This happened in no less than three of the five "known relevant" articles. In the case of the articles by Hart and Soiva (overleaf), because premature rupture is mentioned in the summary of each, the failure to include the term FETAL MEMBRANES is attributed to "indexer omission", while the failure to apply the term to the article by Hazard is attributed rather to lack of exhaustivity of indexing. The depressing fact about this search is that even the best reasonable "hindsight" strategy (i.e., one requiring FETAL MEMBRANES to be present) could not have achieved more than 50% recall.

306 Infantile diarrhea due to E. coli 0:73 or any other unusual serotypes of E. coli.

Recall 2/6 = 33.3% Precision 11/25 = 44%

A recall failure due to a request too specific in relation to the information need. The requester is interested in new enteropathogenic serotypes of E. coli, whether or not there is a specific reference to infantile diarrhea.

479 Quantitative and kinetic aspects of viral antigen-antibody reactions.

Recall 0/5 = 0 Precision 3/24 = 12.5%

This is a highly complex request, and the index language cannot cope with the precise relationships involved. Nevertheless the searcher did not

ABSENCE OF GERMINAL EPITHELIUM WITH NORMAL KARYOTYPE

SIR,—We should like to record the study of the karyotype of a patient with the clinical and histological features of the syndrome characterised by "the complete absence of germinal epithelium, without impairment of the Sertoli or Leydig cells", known also as "germinal cell aplasia".

The patient is a man aged 30 who consulted us for sterility after 6 years of marriage. His past history was negative for diseases affecting the germinal epithelium. He is 5 ft. 10 in. in height and his intelligence, general physique, secondary sexual characteristics, and phallus are normal. The testes are small (2 × 2 cm.) and soft.

The excretion of 17-ketosteroids if 6·6 mg. per 24 hr. The urinary gonadotrophins are +6 and −52 units per 24 hours, and in tests on several occasions the patient had azoospermia.

Testicular biopsy (fig. 1) showed great decrease in size of the seminiferous tubules and absence of germinal epithelium. Inside the tubules only Sertoli cells were found. The number of Leydig cells was increased. There were no Barr bodies in the buccal smears, and no drumsticks in the leucocytes.

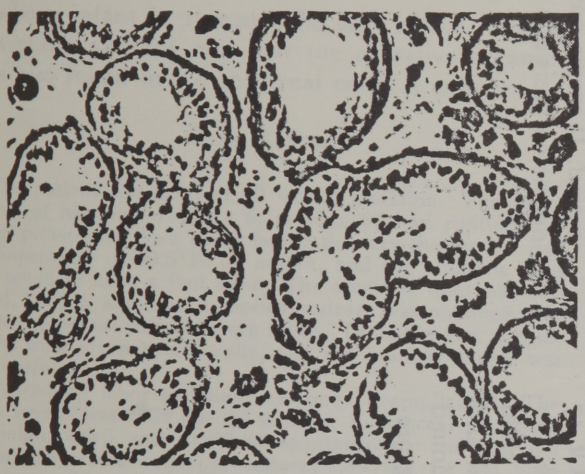


Fig. 1—Testicular biopsy (×100).

We found 46 chromosomes in 42 metaphases in short-term tissue cultures of connective tissue and skin, and a normal male karyotype in 8 cells (fig. 2). So far as we know, chromosome studies have not been previously reported in typical cases of this syndrome, but many of the cases known as "negative-Klinefelter" are regarded by de la Balze et al. as different grades of the syndrome of "germinal aplasia". These chromatin-negative cases of Klinefelter's syndrome are known to have an XY karyotype.

Quite possibly other conditions associated with defective spermatogenesis, without known cause, may represent various stages in a process whose extreme manifestation is the syndrome of "germinal aplasia". Klotz et al. found an XY karyotype in 4 patients with abnormal spermatogenesis and in another with a typical syndrome of "germinal aplasia" he found negative chromatin but no chromosomal analysis was made.

The normal male karyotype in all these patients and in the one here described suggests that "germinal aplasia syndrome", "negative-Klinefelter", and arrested or abnormal spermatogenesis are different manifestations of a gene mutation and not of a chromosomal aberration. Some support for this hypothesis can be found in similar histological abnormalities and sterility in the mouse caused by gene mutations. Mintz[1] showed in homozygous embryos of the mouse that in the case of one or both of the W or S1 mutant genes, the number of primordial germ-cells does not increase, owing to mitotic failure during the period of migration, and that this results in sterility in the adult animals. Mintz showed also that at least one normal allele in each one of these two independent loci must be present to obtain normal gametogenesis. Neither the W gene nor the S1 gene alone can induce the proliferation of normal primordial germ-cells. In the homozygotes for one of the two mutant W or S1 genes, the grade of arrest of spermatogenesis is dependent upon the presence of modifiers. Homozygotes for either S1 or W result in a complete absence of germ-cells in some strains of mice, but in other strains with different genetic background they produce only abnormal or arrested spermatogenesis.

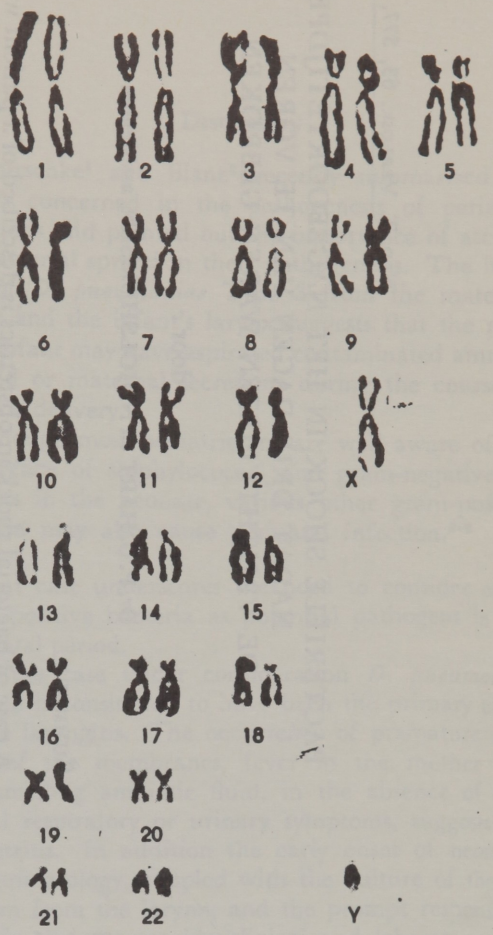


Fig. 2—Karyotype of patient with absence of germinal epithelium.

We suggest, therefore, that in man some of the cases of sterility with a normal karyotype, such as the negative Klinefelter syndrome, the germinal-aplasia syndrome, and some types of abnormal or arrested spermatogenesis, could be due to mutant genes, acting in the way described by Mintz in the mouse.

This work was supported in part by a grant from the Argentine National Research Council.

J. I. VALENCIA
C. B. DE LOZZIO
R. MORENO.

Laboratory of Genetics, Faculty of Sciences, University of Buenos Aires, Argentina.

F. A. DE LA BALZE.

Institute of Endocrinology, Argentine National Institute of Health, Haedo, Buenos Aires, Argentina.

CONSERVATIVE TREATMENT OF THREATENED PREMATURE LABOUR

BY

K. SOIVA AND O. CASTRÉN

SUMMARY

Among 144 patients with a duration of pregnancy of 25 to 34 weeks who were admitted to hospital because of threatening premature delivery (pains, hæmorrhage and/or rupture of the membranes), and were not delivered within 24 hours of admission, conservative therapy was successful in postponing delivery for at least two weeks in 30 per cent. A living child who survived the neonatal period was born in 74 per cent of the cases. Of 47 patients who were given allylestrenol as a routine measure, delivery could be postponed in more than half, and a living child was obtained in four fifths of the cases. Since the neonatal and late prognosis of small-size premature infants is poor, a continued search for possible means to prevent and arrest threatening premature labour should be made.

NED. T. VERLOSK. 63, 377, 1963

BACTERIËLE SHOCK IN HET NAGEBOORTETIJDPERK NADAT DRIE DAGEN VAN TE VOREN DE VRUCHTVLIEZEN WAREN GEBROKEN

door

Dr. P. G. HART en N. HOLSHUIJSEN, assistenten

SUMMARY

A case of bacterial shock in obstetrics is described of a patient whose fetal membranes were ruptured three days before labour started. The renal functiondisturbances of this syndrome are specially discussed. It seems possible to explain the various aspects of this syndrome on basis of a generalized SHWARTZMAN reaction.

PNEUMOCOCCAL LARYNGITIS IN THE NEWBORN INFANT*

Report of a Case

GERALD [illegible],† PHILIP [illegible] M.D.,‡ AND DAVID [illegible] M.D.§

Introduction

LARYNGITIS is [illegible] in the newborn infant, and most [illegible] of pediatrics do not even mention this subject. Disturbances of the larynx in the perinatal period usually result from trauma or congenital defect. The following case of laryngitis due to *Diplococcus pneumoniae* Type 3 in a newborn infant is presented because of the rarity of this entity and the fact that the causative organism was also found in the maternal cervical culture.

Case Report

The patient was born at 40 weeks' gestation. The membranes of his mother (para 6, gravida 6) had ruptured 23 hours previously. Two hours before delivery the mother had a temperature of 99.6°F. by mouth, and the amniotic fluid was noted to have a foul odor. The amniotic fluid was cultured, and, in addition, a cervical culture was obtained. No further maternal cultures were taken. Penicillin and streptomycin were given parenterally, and the hospital course was uneventful.

The 2-hour labor and delivery were uncomplicated. The birth weight was 3005 gm. (6 pounds, 10 ounces). At birth the infant cried and breathed spontaneously and did not require any resuscitative procedure; however, he did have transient mild respiratory distress characterized by generalized rhonchi, slightly decreased air exchange and minimal peripheral cyanosis. The remainder of the physical examination, including the cry and reflexes, was within normal limits.

At 12 hours of age a hoarse cry developed, progressing over the next 72 hours to aphonia. On the 3d day of life the patient became febrile, with the highest recorded temperature of 101.6°F. by rectum. Direct laryngoscopic examination revealed erythematous, edematous vocal cords. No other abnormalities were noted on this examination. Physical examination was unremarkable except for the aphonia. The total white-cell count was 18,000, with 45 per cent neutrophils and 55 per cent lymphocytes. X-ray study of the chest gave no evidence of cardiac or pulmonary disease. Cerebrospinal-fluid examination was normal. Blood and cerebrospinal-fluid cultures were sterile. On the 3d day the culture reports on the amniotic fluid and cervix of the mother revealed that *D. pneumoniae* Type 3 had been present. The direct culture of the baby's larynx subsequently grew out the same organism.

Aqueous crystalline penicillin G, 100,000 units intramuscularly every 6 hours, was started. The temperature returned to normal within 6 hours, and the voice became audible 24 hours after initiation of therapy. The patient received 7 days of antibiotic therapy and had an uneventful hospital course.

Discussion

Benirschke[1] and Blanc[2] recently summarized the factors concerned in the pathogenesis of perinatal infections and pointed out the importance of ascending bacterial spread in their pathogenesis. The isolation of *D. pneumoniae* Type 3 from the maternal cervix and the infant's larynx suggests that the newborn infant may have aspirated contaminated amniotic fluid or maternal secretions during the course of labor or delivery.

Although most pediatricians are well aware of the occurrence of staphylococcal and gram-negative infections in the neonate, various other gram-positive bacteria may also cause perinatal infection.[3-6] The present case underscores the need to consider other gram-positive bacteria as potential pathogens in the perinatal period.

In the case under consideration *D. pneumoniae* Type 3 is considered to have been the primary cause of the laryngitis. The occurrence of premature rupture of the membranes, fever in the mother and foul-smelling amniotic fluid, in the absence of maternal respiratory or urinary symptoms, suggests endometritis. In addition the early onset of neonatal symptomatology, coupled with the culture of the organism from the larynx, and the prompt response to specific therapy provide clinical and laboratory support for the probability that this was a primary pneumococcal infection of the larynx.

Aphonia is abnormal in the newborn infant and should always be evaluated. When it occurs in association with fever and premature rupture of membranes a complete bacteriologic and possibly virologic investigation should be undertaken. By investigating the aphonia in the present case we were able to determine the specific etiologic agent involved and the probable sequence of events.

Summary

A case of neonatal pneumococcal laryngitis is reported. The organism, *Diplococcus pneumoniae* Type 3, was isolated from the amniotic fluid, the maternal cervix and the newborn infant's larynx. The infant was treated with aqueous penicillin and made an uneventful recovery. This was thought to be the first case report of pneumococcal laryngitis in a newborn infant.

exhaust all reasonable approaches to retrieval. In fact, the fairly simple-minded strategy below would have retrieved three of the five relevant articles (60% recall).

[Virus or virus disease terms] AND [ANTIBODIES / ANTIGENS / ANTIGEN-ANTIBODY REACTIONS] AND [quantitative test terms or KINETICS]

It should have been capable of retrieving all five. However, the term KINETICS was not applied to two articles dealing with the kinetics of interaction between virus and antibody; in one article the term "kinetics" appears in the title.

506 Treatment of finger tip amputations in children and adults.

Recall 3/9 = 33.3% Precision 14/20 = 70%

A poor search formulation was responsible for the failure of this search. Although the requester asks for treatment (i.e., repair) of finger tip amputations, the searcher made no use of any treatment terms, relying solely on the coordination of "finger" terms and "amputation" terms. Since by no means all articles on damaged finger tips are indexed under AMPUTATION (the term appears to be used more for therapeutic than for traumatic amputation) or AMPUTATION STUMP, much of the relevant material was missed. Most could have been retrieved on FINGER INJURIES and SKIN TRANSPLANTATION.

527 Normal and abnormal development or growth of the eye, particularly the lens.

Recall 2/9 = 22.2% Precision 11/25 = 44%

The searcher did not cover all reasonable approaches to the retrieval of literature on "normal and abnormal growth". LENS, CRYSTALLINE and [EMBRYO / CHICK EMBRYO] were combinations not used. LENS, CRYSTALLINE and REGENERATION would have retrieved additional relevant articles on normal growth, and CATARACT/ETIOLOGY would have retrieved further articles on abnormal growth. The inclusion of the above terms could have raised retrieval to 66%. Other recall losses occurred through indexer omissions.

Having re-examined ten "bad" MEDLARS searches, and ten "good" ones, and drawing upon all the analyses previously presented in this report, we can now enumerate the factors that most critically affect the performance of a MEDLARS search:

1. The prime factor is the quality of the interaction between the

requester and the system. Given a request statement that inadequately represents the information requirement, there is nothing that a searcher can do (except purely by chance) to produce a good search result. Obviously, the situation is most serious when the request statement is more specific than the actual information need.

2. Given a request statement that closely matches the requester's needs, one factor that will influence the search results is the complexity of the request. The "simpler" the request (i.e., the fewer facets involved), the better the result is likely to be.* A search that requests virtually everything on syringomyelia is single-faceted, and involves only one MeSH term. With such a broad request, given appropriate index terms, it should be possible to obtain high recall and high precision. Since the requester has general needs, he will tend to accept as relevant any article that bears in some significant way on the subject of syringomyelia. However, consider a request on "roentgenologic joint changes in syringomyelia". This is a more complex request, involving three facets (the disease facet, the diagnostic technics facet, and the anatomic facet). Many more MeSH terms are involved, and the relationship between these terms becomes important. Moreover, the requester's relevance standards will be much more stringent - he is unlikely to accept as relevant any article that does not discuss the precise topic of roentgenologic changes in syringomyelia. Consider a third request: "spontaneous dislocation of the atlas simulating syringomyelia". This is much more complex, involving exact relationships between index terms, and the requester is likely to be very stringent in his requirements. With this type of request we are liable to get both false coordinations and incorrect term relationships. It is possible to get high recall on any of these three searches, but this recall will be achieved with a precision that is likely to decrease dramatically with the complexity of the request.

3. Obviously, the MEDLARS performance for any request will depend upon the ability of the index language to precisely express the notions involved. We can do a good search on experimental renal hypertension because specific terms are available to cover this topic. On the other hand, a search on gallbladder perforation was a failure because there is no specific term to cover "perforation".

If there is no specific MeSH term for a particular notion, but the entry vocabulary indicates which more general term or term combination is to be used to express the topic, precision failures will occur, but recall should be unaffected. If there is neither a MeSH term nor an entry vocabulary term for a particular notion (as in the case of "premature rupture of the fetal membranes"), we are likely to get both recall failures

* In the earlier discussion of searching using term weighting, it was noted that the average recall ratio for the 16 searches selected as involving single key notions was 74.5% while the average precision ratio was 48.6%. This performance is substantially better than the average performance over all 299 test searches.

and precision failures, because:

(a) The topic is susceptible to omission in indexing, because the indexer does not know how to express it.

(b) Where expressed, there is likely to be inconsistency in how it is expressed.

(c) The searcher does not know how the topic has been indexed.

4. Related to both complexity of request, and adequacy of the vocabulary, is the matter of the subject field of the request. Table 21 presents a breakdown of performance figures by subject field. Although no general, consistent pattern emerges from these figures, we can nevertheless discern some trends. The DRUG/DISEASE requests (drug therapy) achieved the lowest recall ratio, but the highest precision, of any group. This is due to the fact that the MEDLARS indexing has not been sufficiently exhaustive to cope with the type of request that asks for several medicinal agents in various combinations or for a comprehensive search on all applications of a particular drug.

Taking both recall and precision into account, the DRUG/BIOLOGY (pharmacology) and the BEHAVIORAL SCIENCE requests perform noticeably worse than requests in other subject areas. In the case of the BEHAVIORAL SCIENCE requests, not only is the language of the subject field somewhat imprecise, but also analyses have shown the MeSH terminology to be weak in this area. One of the major problems with the DRUG/BIOLOGY requests was the tendency for incorrect term relationships to occur. In particular, before the introduction of subheadings, it was difficult to distinguish therapeutic use of a drug from adverse effects.

The PHYSICS/BIOLOGY requests are mostly related to radiation effects. These searches tend to achieve high recall, because they deal with fairly tangible subject matter, and general radiation terms exist in the system. However, they tend not to achieve high precision because the vocabulary does not completely differentiate all types of radiation (e.g., ionizing from non-ionizing, masers from MICROWAVES in general).

The requests in the area of TECHNICS are usually fairly tangible (they deal with particular named procedures: surgical, diagnostic, analytical, etc.) and achieve results which, taking both recall and precision into account, are better on the average than the results for searches in any of the other broad subject areas.

These, of course, are only generalizations. The various performance figures indicate trends only. It is the detailed search analysis

that reveals what exactly is happening in the system. As previously mentioned, the DISEASE searches will perform well or badly depending upon the existence of appropriate specific terms. Likewise a BEHAVIORAL SCIENCE search on a topic for which specific terms exist will perform better than a TECHNICS search on a topic for which no specific terms exist.

In relation to subject area, the most meaningful analyses are those (presented earlier in Tables 11, 12 and 13) that correlate subject with lack of specificity in the vocabulary and with the propensity for false coordinations and incorrect relationships between terms.

Moreover, these subject fields are very broad and hide specific problem areas that were brought to light in the search analyses. One of these areas is that of epidemiology. There were six test searches relating to epidemiology (#30, 64, 65, 71, 77, 117). These, on the average, achieved 62.1% recall but only at 39.5% precision (very low in relation to the average for all MEDLARS searches). The problem here is that certain terms (such as STATISTICS and NEOPLASM STATISTICS), necessary to retrieve material indexed prior to the introduction of the subheading OCCURRENCE, have not been used with a strictly epidemiologic connotation. They have also been applied to articles discussing, for example, success rates for various therapeutic procedures. Consequently, although it is possible to obtain reasonable recall in these searches, this can only be achieved at a comparatively low precision.

A more acute problem area is that of immunology. There are nine searches (#10, 43, 44, 118, 162, 215, 471, 479, 569) in this field, and they average only 46.8% recall at 49.1% precision. This is a difficult subject, with complex requests, that presents problems to both indexers and searchers. Where such a problem area is known to exist, it seems essential that special efforts be made to (a) instruct indexers and searchers in the elements of the subject, and (b) clarify MeSH terminology by careful definitions.

5. Indexing policies and practice will control the performance level of a search, as demonstrated clearly by the search on premature rupture of the fetal membranes (# 177) and that on testicular biopsy (# 174). In the former, even the best "hindsight" search formulation could only achieve 50% recall because of the omission of the term FETAL MEMBRANES from the indexing of several relevant articles. As previously mentioned, such omissions are most likely to occur in situations in which no specific terms exist in the vocabulary. Similarly, a high recall on testicular biopsy was impossible because the fact that a biopsy was conducted is not brought out routinely in indexing.

6. Finally, given a request that matches the information need, given appropriate specific terms in the vocabulary, and given adequate indexing, a search can be completely ruined (as in # 174 and # 506 above) or substantially reduced in value by an inadequate search formulation.

Table 21

Breakdown of performance figures by subject field

Subject field	Number of searches	Precision ratio	Recall ratio
DISEASE	110	48.1%	59.7%
PRECLINICAL SCIENCES	85	53.7%	59.0%
TECHNICS	58	53.7%	63.4%
DRUG/BIOLOGY	27	43.1%	51.2%
BEHAVIORAL SCIENCE	17	51.2%	54.2%
DRUG/DISEASE	14	60.2%	41.8%
PHYSICS/BIOLOGY	12	45.4%	63.2%
PUBLIC HEALTH	6	44.8%	51.8%

VARIATIONS IN PERFORMANCE BETWEEN FIVE MEDLARS CENTERS

The test searches were formulated at five MEDLARS centers, thus allowing some analysis of performance variations between these various centers. Table 22 presents a breakdown of performance figures, for the 299 searches, by the MEDLARS center formulating the search strategy. This is an extremely interesting table. It indicates not that certain centers are clearly superior to others but that there are significant differences between centers in policies and tactics. If we rank the centers by recall performance, we get an order that is the complete inverse of the order we get when we rank the centers by precision performance:

Recall	Precision
UCLA	COLORADO
HARVARD	NIH
NLM	NLM
NIH	HARVARD
COLORADO	UCLA

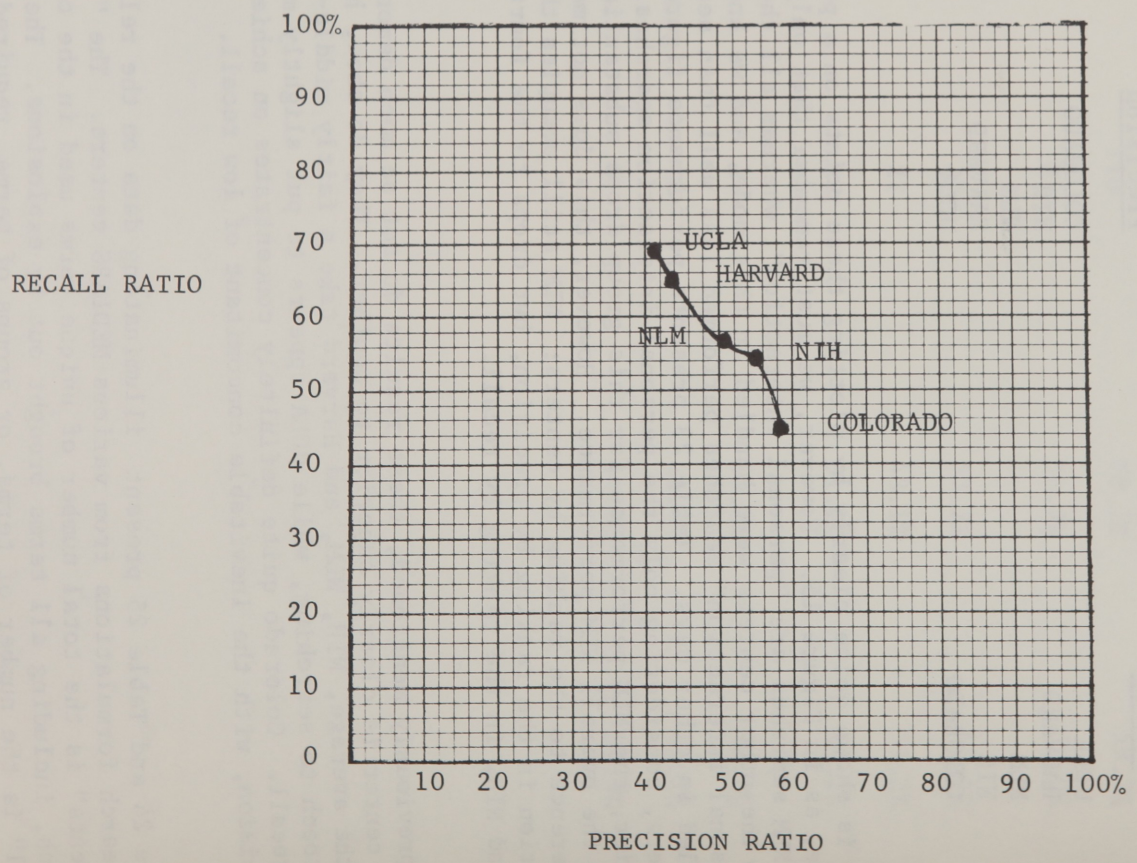
This is shown quite clearly by plotting these points on a performance curve, as in Figure 15. However, we must remember that only a few of the NLM searches are "personal interaction", whereas all the searches from the other centers were handled in this mode, and we know that the "personal interaction" searches performed less well than searches handled in other ways. Table 23 presents performance figures for the centers, considering only the personal interaction searches. As noted earlier, the NLM performance for this group drops substantially lower than the overall NLM performance. However, this does not make too much difference to the ranking by centers: NLM still occupies the middle position in the ranking by precision, but drops to the fourth position, behind NIH, in the ranking by recall.

As previously mentioned, these results do not so much demonstrate that one center is clearly superior to another. They do show, however, that, on the average, NIH, NLM, and Harvard take a fairly middle-of-the-road approach to searching, while UCLA appears to put slightly more emphasis on recall. Colorado quite definitely concentrates on achieving a high precision, with the inevitable concomitant of low recall.

Table 24 and Table 25 present illuminating data on the relative complexity of search formulations from various MEDLARS centers. The "number of search elements" is the total number of unique terms used in the conduct of a search, including all terms brought out by explosions. The "coordination level" is the number of terms, or groups of terms, required to co-occur to

Figure 15

Recall/precision plot showing performance points for various MEDLARS centers



cause retrieval of an article (i.e., it is the number of coordinates in a logical product relation). LENS, CRYSTALLINE and VERTEBRATES (explosion) is a two-term coordination. LENS, CRYSTALLINE and VERTEBRATES (explosion) and HISTAMINE is a three-term coordination. In this analysis, the coordination level is that on which the searcher largely relied. For example, if a searcher used one 3-term coordination in a total of, say, 50 term coordinations, but all the rest were 2-term, this would be counted as a 2-term coordination.

From these tables (based on all 11 test searches from UCLA and random samples of 20 searches from each of the other centers), we can see great variation in the complexity of search formulations between MEDLARS centers. It can be seen that the search formulations from Colorado and UCLA tend to use many fewer terms than the formulations from NIH, NLM, and, to a lesser extent, Harvard. The "average" is less meaningful, in this analysis, than the median and the mode. Among the 20 Colorado searches, more fell into the range 1-5 terms than any other range. The mode at UCLA was 6-20, while the mode at the other centers was over 100 terms per formulation.

Colorado relied largely on a coordination level of 2 in the test searches. None of the Colorado searches was conducted primarily at a coordination level greater than 2-term. The other centers made some use of higher coordination levels. NLM made greater use of these higher coordination levels than the other centers; approximately 35% of the NLM searches involved the coordination of more than two terms.

It is also interesting to consider variations in the number of citations retrieved per search. These data, based on 299 searches, are given in Table 26. Over 28 searches, Colorado retrieved an average of 154 citations per search, while NLM averaged 160 citations over the 198 searches; Harvard, UCLA and NIH achieved higher average retrievals with 199, 249, and 260 respectively. Again, averages are misleading in this context. For example, Colorado normally retrieves few citations, but their average is dominated by a few very broad searches (e.g, # 9 on the single term HYPOTHERMIA, INDUCED which retrieved 860 citations, and # 98 on vagotomy, which retrieved 875). Moreover, the 500 citation ceiling on searches run at NLM affected the average retrieval for all centers except Colorado, because all test searches from NIH, UCLA and Harvard were actually run at NLM. The "median" retrieval shows Colorado and, to a lesser extent, NLM, as tending to retrieve fewer citations than the other centers. The mode shows UCLA as having a penchant for very large retrievals, which accounts for their comparatively high recall, low precision performance.

Of course, when we compare the performance of various centers, we have to remember that indexing and index language are constants. The only things that can vary between centers are the searching strategies and

Table 22

Breakdown of performance figures by MEDLARS center formulating search strategy

Center	Number of searches	Precision ratio	Recall ratio
UCLA	11	40.7%	69.2%
COLORADO	28	57.2%	43.3%
NIH	21	55.6%*	55.5%*
HARVARD	41	43.2%	64.6%
NLM	198	50.9%	57.9%

* The figures for NIH should really be considered as "NIH revised by NLM", since these searches were processed at a time when NLM was substantially revising searches from the centers. Four of these NIH searches were virtually completely reformulated at NLM. If these four are omitted from the calculation, the NIH figures become 54.5% precision and 59.3% recall.

Table 23

Breakdown of performance figures by MEDLARS center, considering only "personal interaction" searches

Center	Number of searches	Precision ratio	Recall ratio
UCLA	11	40.7%	69.2%
COLORADO	28	57.2%	43.3%
NIH	21	55.6%*	55.5%
HARVARD	41	43.2%	64.6%
NLM	8	48.3%	45.2%

* "NIH as revised by NLM". See footnote to preceding table.

Table 24

Complexity of search strategies at various MEDLARS centers

	Average number of search elements	Range	Median	Mode*
COLORADO	50.8	2-425	10.5	1-5
NLM	152	4-767	81	over 100
UCLA	85.9	6-437	38	6-20
NIH **	171.7	1-568	85.5	over 100
HARVARD	94.4	1-416	50	over 100

*On the scale 1-5, 6-20, 21-30, 31-40, 41-50, 51-100, over 100.

**Eliminating the three searches that were substantially altered at NLM, the NIH figures become as follows:

Average number of search elements: 143.2
Range: 1-421
Median: 74
Mode: over 100

Table 25

Coordination levels used in search strategies at various MEDLARS centers

Colorado:	100% of searches were conducted at a coordination level of 2 or 1.
NLM	65% of searches were conducted at a coordination level of 2.
UCLA:	82% of searches were conducted at a coordination level of 2.
NIH:	85% of searches were conducted at a coordination level of 2 or 1.
Harvard:	85% of searches were conducted at a coordination level of 2 or 1.

Table 26

Numbers of citations retrieved by various MEDLARS centers

	Average	Range	Median	Mode*
Colorado (28 searches)	154	2-875	59	1-50
NLM (198 searches)	160	0-822	92	1-50
UCLA (11 searches)	249	19-533	173	500-550
NIH (21 searches)	260	6-1167	198	1-50
Harvard (41 searches)	199	0-764	171	1-50
OVERALL	175			

*In blocks of fifty.

Table 27

Recall and/or precision failures due to defects of searching strategies.* 302 searches were examined, and 99 found in which this type of failure occurred. Breakdown by MEDLARS center processing search.

MEDLARS center	Number of searches involving failures due to defective strategies	Percentage of total searches involving failures due to defective strategies (99)	Percentage of total searches processed by this center
COLORADO	4	4.0%	14.3%
UCLA	5	5.1%	45.5%
NIH	5	5.1%	23.8%
HARVARD	12	12.1%	29.3%
NLM	73	73.7%	36.9%

* Three types of searching failure are so designated:

(1) Failure to cover all reasonable approaches to the retrieval of relevant literature.

(2) Use of a seemingly inappropriate term.

(3) Use of defective search logic.

Table 28

Recall and/or precision failures due to inadequate user-system interaction. Breakdown by MEDLARS center processing search.

MEDLARS center	Number of searches in which interaction failures occurred	Percentage of total searches processed by center
COLORADO	15	53.6%
NIH	9	42.9%
HARVARD	20	48.8%
NLM	5	62.5%
UCLA	6	54.5%

the user-system interaction. The greatest variability is likely to relate to policies regarding the levels of specificity and exhaustivity adopted by various centers. Tables 27 and 28 present some interesting figures on defective searching strategies, and on the effect of user-system interaction, at the various centers.

A search formulation was regarded as containing a defect if it (a) failed to use all reasonable term combinations in relation to the request, or (b) used a seemingly inappropriate term or term combination, or (c) contained an error in search logic. It can be seen from Table 27 that all centers are prone to defective formulations of one kind or another. Five of the 11 UCLA searches (i.e., 45.5%) involved recall and/or precision failures that were attributed to inadequacies or errors in the search formulation. On the other hand, errors of this type were found in only 4 (14.3%) of the 28 Colorado formulations.

Table 28 shows that inadequate interaction with the requester occurs significantly often at all the centers. In five of the eight "personal interaction" searches conducted at NLM (i.e., 62.5%) it was felt that the "request" as recorded by the search analyst was an imperfect representation of the information need. NIH, with nine out of 21 (i.e., 42.9%) searches affected, has suffered slightly less from inadequate interaction than the other centers.

Of course, this test program has not evaluated the performance of the participating MEDLARS centers on a strictly comparable basis, as would be achieved, for example, by having the same group of test searches formulated at a number of different centers. Nevertheless, general trends are clearly discernible. The search formulations of Colorado are the "simplest" of any. This center employs fewer terms and uses strategies that are predominantly simple two-term coordinations. The terms selected are the few "key" terms in relation to the request, and they tend to be as specific as possible. As expected, this results in a high precision but a low recall. There are few "defective" searching strategies at Colorado because errors are not likely to occur in a simple formulation. The only defect possible is that of failing to use all reasonable approaches to retrieval. The philosophy at Colorado appears to be one of providing an "only-but-not all" (i.e., high precision) search, but this is achieved at the expense of a low average recall. In fact, this philosophy appears to influence the interaction with the requester. Many of the Colorado requests, as recorded by the search analyst, are more specific than the actual area of interest. In other words, the searchers may be over-zealous in interacting with the user, with the effect that the requester is persuaded to accept less than he would really like to see.

UCLA also tends towards fairly simple strategies, although they incline to use more terms than Colorado. However, whereas Colorado chooses specific terms, UCLA tends to search much more broadly (e.g., search # 45, which was broadened from "electrical brain stimulation" to "brain

electrophysiology"). This results in many more citations being retrieved and gives a high recall performance with a comparatively low precision. Nevertheless, if the 40% precision achieved on the average at UCLA is acceptable to the users, then this center performed rather better in the evaluation than the other participating centers.

As stated earlier, the other three centers are more middle-of-the-road, using more complex formulations in an attempt to achieve a balance between recall and precision. Although one tends to think of formulations involving many terms as representing a "shotgun" approach designed to get maximum recall, this is not necessarily the case. The terms may be present to secure high precision rather than high recall. For example, consider search # 194 on various nutritional and toxicological aspects of chromium. Several hundred terms were coordinated with CHROMIUM or CHROMATES in an attempt to cover only the specific aspects of interest.

However, when we consider the complexity of a search formulation we must keep in mind the matter of economics. Presumably a complex strategy takes longer to formulate. It will also involve more computer searching time (although it may save on printing). Sometimes one doubts whether the complexity is really justified. For example, the complex "chromium" strategy was successful in achieving 100% recall (5/5) at 60% precision with a total retrieval of 94 citations. However the single-term strategy CHROMIUM or CHROMATES would have achieved 100% recall and retrieved only about 180 citations, of which approximately one third would be relevant (i.e., a precision ratio of 33%). Perhaps the simple single-term search would have been the more reasonable approach in this case.

In other cases, the reverse occurs, and every possible term is thrown in in an attempt to squeeze out the last ounce of relevant literature. This occurred, for example, in search # 526, on the effect of nephrectomy on the contralateral kidney (i.e., compensatory renal hypertrophy and hyperplasia). Some very general term combinations (e.g., NEPHRECTOMY and FOLLOW-UP STUDIES or POSTOPERATIVE COMPLICATIONS or PROGNOSIS) were included by the searcher. The search retrieved 180 citations, of which only about one third are in any way relevant. Analysis shows that the obvious strategy

		HYPERTROPHY
NEPHRECTOMY	and	HYPERPLASIA
		HYPERTROPHY AND HYPERPLASIA

would have retrieved at least 90% of the major value literature. Moreover, every article indexed under the combinations noted above was judged relevant, so that we could have expected this simple strategy to achieve approximately 90% recall of major value articles at close to 100% precision. Is it worth elaborating, with the intention of getting maximum recall, when appropriate specific terms exist in the vocabulary?

Also related to the economics of searching is the amount of "reformulation" carried out by the searchers. "Reformulation" implies that the original

strategy is run but subsequently abandoned, either because the total number of citations retrieved is unacceptably high or because the search is thought to have brought out an unacceptable amount of irrelevancy, or because of an error discovered in the search logic. None of the Colorado test searches and only two of the Harvard searches were reformulated. One UCLA search and four NIH searches were reformulated at NLM (additional formulations were "revised" at NLM before the search was conducted). Of the 198 test searches conducted by NLM, no fewer than 47 (23.7%) were reformulations. Obviously, reformulation involves additional intellectual effort and thus increases the cost of the search. It also has a pronounced effect on throughput time. For example, the request for search # 27 was received on 9/7/66. Reformulation delayed the search results until 12/23/66. when they were received too late to be of value to the requester.

Performance of individual searchers

There are performance variations between various searchers just as there are performance differences between the various MEDLARS centers. Table 29 presents performance figures for individual NLM search analysts. Taking both recall and precision into account, searchers B, F and I achieved rather better results over 24, 11, and 10 searches respectively than the other searchers. Again, varying policies are evident. Searcher H, for example, achieved the high average recall of 73.7% over 10 searches, but this was accompanied by the comparatively low average precision of 45.5%

Of course, these comparisons show only general tendencies. We cannot legitimately say that one individual performs better than another unless we have the same group of requests formulated by a number of different searchers, and analyze to see how each performs. No effort has been made, at the present time, to determine whether any one of the analysts was given a higher proportion of more difficult searches than the others (for example, search # 177, as already noted, could not reasonably have achieved more than 50% recall for reasons outside the control of the searcher), a fact that would obviously affect the performance level.

307-006 O-68—11

Table 29

Breakdown of performance figures for NLM searches, by individual searcher

Searcher	Number of searches	Precision ratio	Recall ratio
A	80	48.4%	54.7%
B	24	52.2%	61.8%
C	17	49.6%	56.0%
D	16	49.1%	54.0%
E	13	58.2%	56.0%
F	11	64.1%	56.3%
G	10	46.2%	65.1%
H	10	45.5%	73.7%
I	10	52.0%	64.2%
J	4	55.2%	56.2%
K	1	84.2%	71.4%
L	1	42.1%	85.7%
M	1	70.6%	27.3%

MEDLARS INDEXING COVERAGE

From the requester-supplied recall base, we can get some rough idea of the MEDLARS indexing coverage. Collectively the requesters provided a total combined base of 1054 known relevant articles. We have included in this base only documents of the type that MEDLARS includes as a matter of policy; i.e., articles from general scientific journals and journals related to the general area of biomedicine. Excluded from this base were items, named in advance by requesters, that fell into the following categories not normally indexed by MEDLARS: separately published reports; separately published colloquia and proceedings; articles from journals that publish an occasional article of biomedical interest but are generally outside the scope of MEDLARS (e.g., Journal of Heat Transfer and Nucleonics); abstracts; articles predating MEDLARS.

Of the 1054 articles accepted as being within the scope of the system, 940 (89.2%) were found to be in the MEDLARS data base at the time the searches were conducted for the requests concerned. The other 114 articles were not in the data base, at the time of the searches, for the following reasons:

Articles from journals regularly indexed but not in the system at the time of search because of indexing backlogs.	64 (6.1%)
Articles from journals not currently indexed by MEDLARS or not indexed in the year cited (1964).	22 (2.1%)
Articles not indexed for some reason although the containing issue of the journal was indexed.	12 (1.1%)
Articles from journal issues or volumes that "escaped" indexing.	11 (1.0%)
Articles from the "proceedings" section of journals otherwise indexed.	5 (.5%)

From the above we can say that MEDLARS was found to include 89.2% of the relevant journal literature, within the scope of the system, as known to requesters at the time they made their requests. Assuming that the 64 articles in the indexing backlog would eventually get into the system, we can say that the ultimate MEDLARS coverage of the relevant journal literature, as known by requesters, was 95%.

Seventeen separate journals, of a general scientific nature or with some biomedical orientation, not indexed by MEDLARS or not indexed in

the year cited, were cited by requesters. Most of these were single citations.

About 1% of all the articles cited by requesters were found not to have been indexed, although the journal issue containing the article had been indexed. Two of these were letters to Lancet, two were editorials, some were from journals selectively indexed only, and others were not indexed for no apparent reason.

Another 1% of all articles cited were found to be contained in journal volumes or journal parts that for some reason had escaped indexing. For example, single issues of Nature and Science were found not to have been indexed, two whole volumes of Biochemical and Biophysical Research Communications were never received in the Index Section, and a whole year of Mutation Research was not even received by the Library (as of 12/31/66 nothing from this journal had been indexed since 1964).

Other articles, making up about 0.5% of all those cited by requesters, were not indexed because they appear in the "proceedings" section of journals such as the Journal of Physiology and the Biochemical Journal. MEDLARS does not normally index such conference proceedings.

FOREIGN LITERATURE USAGE FACTORS

Of the 302 requesters for whom test searches were completed, 55 (18.2%) requested that only English material be retrieved. An additional 35 searches (11.6% of the total) were conducted with some language restriction (e.g., "English, French and German") imposed. There were no language restrictions placed on the remaining 212 searches (70.2% of the total).

Table 30 presents data for the 35 searches in which a partial language restriction was imposed. The combined random sample (i.e., the articles derived by random sampling from each search and submitted to the requester for his assessment) for the 35 searches totalled 768 articles, of which 646 (84%) were in English and 122 (16%) in other languages.

Of the 122 foreign articles retrieved, the requesters were able to assess 97 (79.5%). Of the assessed items, 42 (43.3%) were judged of value. In the case of twelve of the 25 unassessed items (48%), the requester intended to take some further steps (e.g., getting at least a partial translation or attempting to find an English abstract) to determine their value.

Table 31 presents similar data for the 212 searches in which no language restriction was imposed. The combined random sample for these searches totalled 4627 items, of which 3242 (70.1%) were in English and 1385 (29.9%) in other languages. Of the 1385 foreign articles, 867 (62.6%) were assessed, and 346 (39.9%) of the assessed items were judged relevant. In the case of 127 (24.5%) of the unassessed items, the requester intended to take some further action to determine their value.

For the searches in which no language restrictions are placed, MEDLARS retrieves, on the average, in the proportion of 70% English material to 30% foreign. Note the discrepancy between this proportion and the estimated proportion of English articles to non-English in the total data base (55% to 45%). This discrepancy is due to the fact that the percentage of the English material indexed in depth is much higher than the percentage of the foreign material indexed in depth. Obviously, all other things being equal, if we have two sets of articles, one indexed at a higher exhaustivity level than the other, the exhaustively indexed portion of the file will account for proportionately more of the total retrievals than the non-exhaustively indexed portion.

It is interesting to note the variations between Table 30 and Table 31. In the 35 searches summarized in Table 30 the requester has indicated the languages in which he is willing to accept material. This would tend to indicate that these are the languages with which the requester is able to cope. Nevertheless, these 35 requesters were only able to assess 79.5% of the total foreign articles retrieved. This compares with an assessment rate of 62.6% for the foreign material retrieved in the 212 searches in which there were no language restrictions.

Table 30

Language	Total in samples	Assessed		Relevant		Unassessed items for which further action is to be taken	
		Number	Percentage of retrieved and sampled	Number	Percentage of assessed	Number	Percentage of unassessed
GERMAN	65	51	78.5%	18	35.3%	6	42.9%
FRENCH	38	36	94.7%	21	58.3%	1	50.0%
SPANISH	7	4	57.1%	0	0	2	66.7%
ITALIAN	6	6	100.0%	3	50.0%	-	-
JAPANESE	3	0	0	-	-	2	66.7%
RUSSIAN	2	0	0	-	-	0	0
POLISH	1	0	0	-	-	1	100.0%
TOTALS	122	97	79.5%	42	43.3%	12	48.0%

Table 31

Language	Total in samples	Assessed		Relevant		Unassessed items for which further action is to be taken	
		Number	Percentage of retrieved and sampled	Number	Percentage of assessed	Number	Percentage of unassessed
GERMAN	342	231	67.5%	98	42.4%	32	28.8%
FRENCH	313	250	79.9%	106	42.4%	19	30.2%
ITALIAN	183	109	59.6%	42	38.5%	17	23.0%
RUSSIAN	150	63	42.0%	19	30.2%	18	20.7%
JAPANESE	90	37	41.1%	9	24.3%	16	30.2%
SPANISH	72	46	63.9%	18	39.1%	3	11.5%
POLISH	66	38	57.6%	20	52.6%	6	21.4%
CZECH	47	26	55.3%	12	46.2%	4	19.0%
HUNGARIAN	26	10	38.5%	0	0	2	12.5%
DUTCH	21	12	57.1%	4	33.3%	4	44.4%
PORTUGUESE	18	13	72.2%	5	38.5%	1	20.0%
RUMANIAN	8	6	75.0%	2	33.3%	1	50.0%
HEBREW	7	4	57.1%	2	50.0%	1	33.3%
SWEDISH	7	4	57.1%	3	75.0%	0	0
DANISH	6	5	83.3%	1	20.0%	0	0
SERBIAN	6	3	50.0%	1	33.3%	0	0
BULGARIAN	4	1	25.0%	0	0	1	33.3%
UKRAINIAN	4	3	75.0%	2	66.7%	0	0
FINNISH	4	2	50.0%	1	50.0%	0	0
SLOVAK	3	1	33.3%	0	0	2	100.0%
CHINESE	3	2	66.7%	1	50.0%	0	0
TURKISH	2	0	0	-	-	0	0
KOREAN	1	0	0	-	-	0	0
NORWEGIAN	1	1	100.0%	0	0	-	-
ALBANIAN	1	0	0	-	-	0	0
TOTALS	1385	867	62.6%	346	39.9%	127	24.5%

These 212 requesters were sufficiently interested in only 24.5% of the unassessed articles to take further steps to determine their value, whereas the other group of 35 requesters showed some interest in 48% of the unassessed articles.* The precision ratio for the foreign material in the 212 searches was 39.9%; it was 43.3% in the other group of 35 searches.

Table 31 is of further interest in showing the approximate breakdown of language usage within MEDLARS, and also the ability of requesters to cope with various foreign languages. German accounted for 342 (24.7%) of the total of 1385 foreign articles, French for 313 (22.6%), Italian for 183 (13.2%), Russian for 150 (10.8%), Japanese for 90 (6.5%), Spanish for 72 (5.2%) and Polish for 66 (4.8%). These seven languages appear to account for almost 90% of all the foreign language usage in MEDLARS.

Almost 80% of the French articles were assessable by requesters, as compared with 67.5% of the German articles and 63.9% of the Spanish. Only 42% of the Russian articles were assessed, and only 41% of the Japanese. Polish featured unexpectedly high in the ranking, being seventh in the order by volume of retrieval; 57.6% of the Polish material was assessed and 52.6% of this was judged of value.

We can now give some consideration to the broad question of foreign language usage in MEDLARS. We can expect that approximately 18% of all MEDLARS searches will be conducted on English language material only. An additional 12% approximately will have some more general language restriction, and these will retrieve roughly in the proportion of 84% English to 16% foreign. The bulk of the MEDLARS searches (70%) will be conducted without language restriction and will retrieve in the rough proportion of 70% English to 30% foreign.

Further discussion will be clearer if we put some figures to these proportions and percentages. The 302 test searches retrieved a grand total of 52,570 citations (an average of 174 citations per search). The 55 "English only" searches retrieved a total of 7121 citations (an average of 129 per search). This is as expected. By restricting to "English only" one is reducing the size of the file by about 45%, and one would naturally expect that the average number of citations retrieved would fall, all other things being equal.

The 35 searches with a partial language restriction retrieved 8540 citations (an average of 244 per search). It is difficult to understand why this group of searches should have retrieved, on the average, more

* The decision on whether or not to translate a foreign article correlates strongly with the volume of the published literature on the search topic. If there is a large volume of literature, the requester frequently ignores the foreign material ("sufficient good material in English"). On the other hand, on a more obscure topic, with little literature, the requester is much more likely to have the foreign material translated, at least in part. For example, search # 243, on television ophthalmoscopy, retrieved only one item, a Japanese article and, as expected, the requester indicated that this would be translated.

citations per search than the entire group of 302 searches. The most likely explanation is that these were searches in which a large retrieval was expected, and the search analyst made a special effort to get the requester to restrict the search to languages with which he was fully familiar.

Of the 52,570 citations retrieved in the 302 searches, we can now compute the proportion of English, as follows:

7121 citations from "English only" searches.

84% of the 8540 citations retrieved in the 35 searches with partial language restriction (i.e., 7173 citations)

70% of the 36,909 citations retrieved in the 212 searches without language restriction (i.e., 25,836 citations)

We can now say that, of the 52,570 citations retrieved by the 302 searches, 40,130 (76.3%) are English articles, and 12,440 (23.7%) are foreign.

However, all the English articles can be assessed for relevance, whereas we know that only 79.5% of the foreign articles in the "partial restriction" searches were assessed, and only 62.6% of the foreign articles in the "no language restriction" searches were assessed. Therefore:

79.5% of 1367 foreign articles (i.e., 1087), and

62.6% of 11,073 foreign articles (i.e., 6932) were assessed.

In other words, approximately 8019 of the 12,440 foreign articles were assessed by the requesters.

But only 43.3% of the 1087 articles were judged of value, and only 39.9% of the 6932 articles were judged of value. Therefore, we can say that only about 3237 of all the foreign articles (471 + 2766) retrieved in the 302 searches were of definite value to the requesters.

However, further action to determine relevance was to be taken on 48% of the 280 unassessed items in the 35 "partial restriction" searches (i.e., about 134 articles), and further action was to be taken on 24.5% of the 4141 articles unassessed in the 212 "no language restriction" searches (i.e., 1014 articles approximately). Adding these figures together, we can say quite confidently that the 302 test searches could not have retrieved more than 3237 + 134 + 1014 (i.e., 4385) foreign articles of value to the 302 requesters, and the true figure is probably rather less than this.

However, we know that approximately 50% of all the 52,570 citations were of value, because 50% is the overall precision ratio for these MEDLARS

searches.* Therefore, we can say that the useful foreign language component in these test searches was, at the most, 4385 articles out of 26,285 (i.e., about 16.7%).

Assuming that this group of 302 searches is representative of all MEDLARS searches, as far as the language breakdown goes, we can say that while foreign language articles consume approximately 45% of the total input effort (it takes rather more time, on the average, to index a foreign language article, and foreign language citations will presumably be responsible for proportionately more keypunching problems; to counterbalance this, a higher proportion of the foreign material is non-depth) they do not account for more than 16% of the total usage** in MEDLARS demand searches. On strictly economic grounds of cost-effectiveness, it is difficult to justify 45% of the total effort being expended on input of foreign language material, at least as far as the demand search aspect of MEDLARS is concerned. The usage factor of foreign language articles discovered by manual searching of Index Medicus has not, of course, been determined.

---

* It is interesting to note that, had the 302 test searches all been conducted solely on English material, the average precision ratio might well have been between 5% and 10% higher. The 55 "English only" searches operated at an average precision ratio of 56.3%. The average recall ratio for this group was 57.9%.

** "Usage" relates to articles retrieved and found of value. The 16% figure is an absolute maximum. It is probably highly inflated in relation to the true usage factor, because it was derived on the basis of decisions made on foreign articles "spoonfed" to requesters. Probably a much smaller amount of foreign material is actually requested on the basis of the citation printout only.

Tables 32 and 33 present some very interesting data derived from the combined random sample of articles submitted for assessment in the 302 test searches. This combined random sample consisted of 6491 articles, of which 4884 articles were from depth journals and 1607 from non-depth journals. In other words, about 75% of all MEDLARS retrievals are from the approximately 800 journals now on the "depth indexing" list.* Once more, the discrepancy between this proportion and the proportion of depth to non-depth indexed articles in the base (approximately 55% to 45%) is due simply to the fact that additional index terms are assigned to the depth articles and thus the "depth" portion of the file will receive proportionally greater usage than the non-depth. Note also that the overall precision ratio, calculated by the "average of numbers", for the "depth" articles is 2386/4672 (51.1%), while the overall precision ratio for non-depth is 553/1266 (43.7%).

It is also interesting to see how the retrieval proportion of depth to non-depth varies with the year of citation:

1963 citations: 60.4% $\left(\frac{645}{1067}\right)$ from depth

1964 citations: 73.7% $\left(\frac{1892}{2567}\right)$ from depth

1965 citations: 82.0% $\left(\frac{1676}{2045}\right)$ from depth

1966 citations: 82.5% $\left(\frac{643}{779}\right)$ from depth

This fluctuation corresponds to the increase in the proportion of depth to non-depth in the total file for each of these years (see Table 8):

42:58 in 1964 Index Medicus (i.e., largely 1963 citations)
54:46 in 1965 Index Medicus (i.e., largely 1964 citations)
58:42 in 1966 Index Medicus (i.e., largely 1965 citations)

---

* That is, one third of the journals contribute 75% of the retrievals.

Table 32

Journal usage factors for depth journals

Year of citation	Number of articles judged of value	Number of articles judged of no value	Number of articles unassessed	TOTALS
1962	1	1	0	2
1963	310	285	50	645
1964	878	927	87	1892
1965	823	788	65	1676
1966	356	277	10	643
1967	18	8	0	26
TOTALS	2386	2286	212	4884

Table 33

Journal usage factors for non-depth journals

Year of citation	Number of articles judged of value	Number of articles judged of no value	Number of articles unassessed	TOTALS
1962	2	2	0	4
1963	117	209	96	422
1964	227	304	144	675
1965	145	141	83	369
1966	62	56	18	136
1967	0	1	0	1
TOTALS	553	713	341	1607

Note that there is a levelling-off of the retrieval proportions between 1965 and 1966 citations (an increase of only .5 in favor of the depth articles). This is probably due to the fact, also noted in Table 8, that both depth and non-depth articles are now being indexed with more terms on the average than heretofore, and that the average term gap between "depth" and "non-depth" is showing signs of a gradual diminishing.

However, the fact remains that, at the present time, about 800 depth journals account for approximately 81% of the total MEDLARS usage. This is derived from the data of Tables 32 and 33 which show that depth articles account for 81% $\left(\frac{2386}{2386 + 553}\right)$ of the total articles retrieved and judged of value. As with the foreign language material, we must conclude that the 19% usage factor for non-depth articles is small in relation to the proportion (45%) of the total data base occupied by citations from non-depth journals.* As previously mentioned, in discussing system failures due to indexing, a policy of treating each article on its own merit, whatever the journal it comes from, might be a more sensible approach than the present policy of dividing journals into "depth" and "non-depth".

If the present distinction between depth and non-depth journals is maintained, it might well prove advantageous to split the search file on this basis. The non-depth articles would not be searched routinely for every request, but only for those in which the requester insisted on maximum recall or those on topics likely to be covered primarily in non-depth journals (e.g., nursing and hospital administration). The non-depth file, being indexed with fewer terms and more general terms, might require slightly different searching strategies for optimal use.

The 6491 articles in the combined random sample were drawn from 1387 separate journals (652 depth and 735 non-depth). Of these 1387, approximately 140 are titles no longer indexed by MEDLARS. In other words, over 6491 retrievals, approximately half of the 2400 journals currently indexed were not represented by even a single citation. This re-inforces the impression that a very small proportion of the total journals indexed is responsible for a very high proportion of the total MEDLARS retrievals.

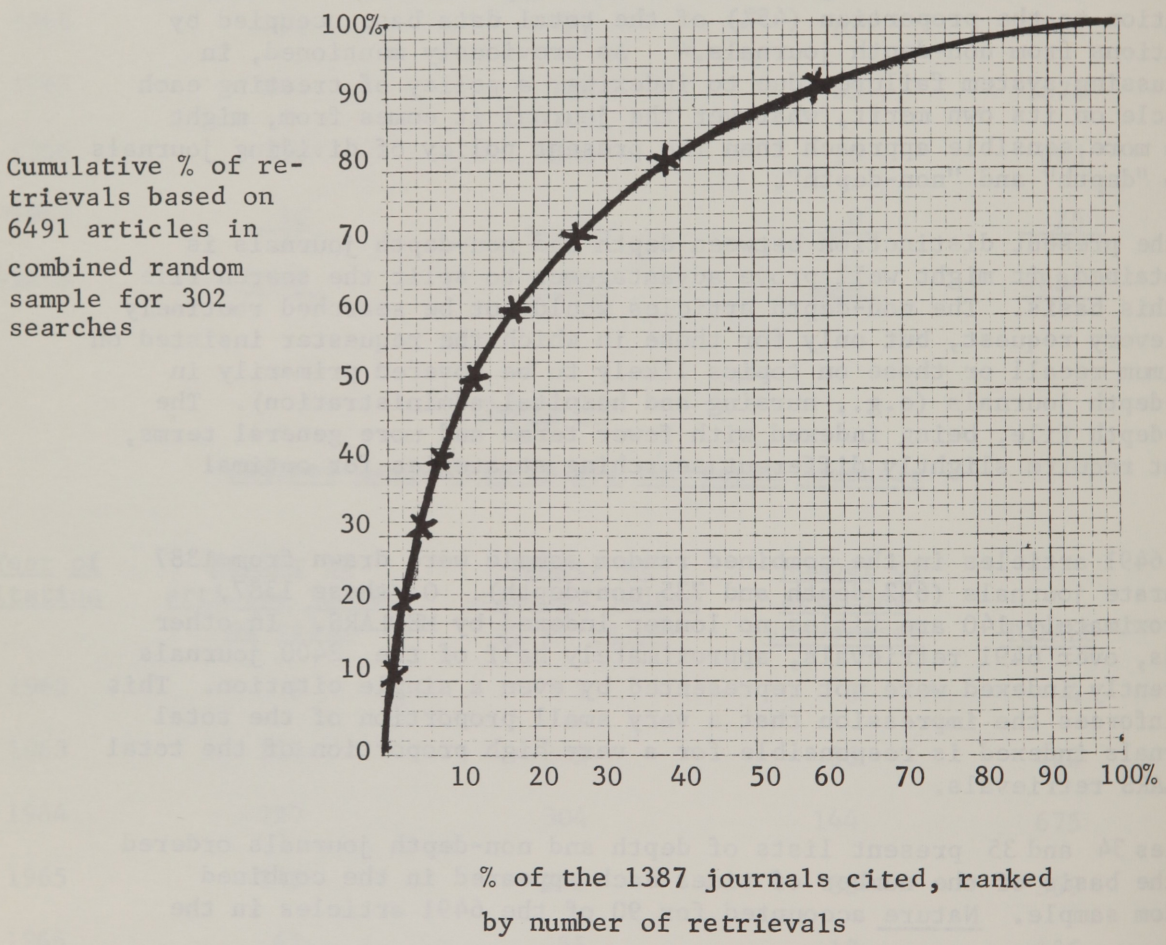
Tables 34 and 35 present lists of depth and non-depth journals ordered on the basis of the number of times each appeared in the combined random sample. Nature accounted for 90 of the 6491 articles in the

---

* As previously noted, a smaller percentage of the retrieved non-depth citations are deemed of value (43.7%) than the percentage of depth judged of value (51.1%). A spot check on 202 non-depth articles, retrieved and judged relevant, showed that 84 (41.6%) were of major value.

Figure 16

MEDLARS retrievals plotted against journals contributing to these retrievals.



combined random sample, and thus tops the list. Practitioner was the non-depth journal that appeared most often, but it only accounted for 19 articles in total.

By combining the two ranked tables, we can determine what percentage of the 1387 journals accounted for what percentage of the total retrievals, based on the combined random sample of 6491 articles. The results are as follows:

The top 10 journals (0.7% of 1387) accounted for approximately 10% of the retrievals (662 articles)

31 journals (2.2%) accounted for 20% of the retrievals (1302 articles)

60 journals (4.3%) accounted for 30% of the retrievals (1954 articles)

98 journals (7.1%) accounted for 40% of the retrievals (2601 articles)

157 journals (11.3%) accounted for 50% of the retrievals (3253 articles)

238 journals (17.2%) accounted for 60% of the retrievals (3894 articles)

356 journals (25.7%) accounted for 70% of the retrievals (4544 articles)

532 journals (38.3%) accounted for 80% of the retrievals (5193 articles)

814 journals (58.7%) accounted for 90% of the retrievals (5841 articles)

1387 journals (100%) accounted for 100% of the retrievals (6491 articles)

These data, when plotted, produce the curve shown in Figure 16. It can be seen that 20% of the journals cited in the combined random sample accounted for over 60% of the retrievals, while 30% of the journals accounted for almost 75% of the retrievals. We hypothesize that this curve will approximate to the distribution of the usage of all MEDLARS journals in all the MEDLARS demand searches. That is, a comparatively small number of all the journals indexed will account for a large percentage of the total demand search retrievals.

One last analysis can be conducted from the data of Table 32 and Table 33.

When we calculate the precision ratio, by the average of numbers, based on year of citation, we arrive at the following interesting pattern:

1963 citations: $\frac{427}{921} = 46.4\%$

1964 citations: $\frac{1105}{2336} = 47.3\%$

1965 citations: $\frac{968}{1897} = 51.0\%$

1966 citations: $\frac{418}{751} = 55.7\%$

1967 citations: $\frac{18}{27} = 66.7\%$

which shows that the precision ratio increases with the recency of citation. One could say, of course, that, on the whole, the more recent material is more likely to be judged of value than the earlier material. However, this is unlikely to be the explanation of the above figures because this would imply a value judgement on the part of the requester, and we know that comparatively few articles were rejected on pure value grounds.

These figures must partly be attributed to the proportion of depth to non-depth articles retrieved which, as we previously mentioned, increases with the recency of the citation. Since the proportion of depth judged relevant is higher than the proportion of non-depth, on the average, these changing retrieval proportions will obviously affect the overall precision ratio.

A second contributing factor is the gradual increase in specificity of the vocabulary over the years. We have already noted that many of the precision failures attributed to lack of specificity were due to the indexing of the earlier material, when fewer specific terms were available. As the vocabulary becomes more specific, higher precision becomes possible. The introduction of subheadings in 1966 increased the specificity of the vocabulary considerably, and also significantly reduced the likelihood of precision failures due to false coordinations and incorrect term relationships.

Also, as we know from previous analyses, the indexing depth has been increasing over the years. The fact that more terms are applied, on the average, allows the subject matter of articles to be indexed more specifically.

Table 34

Ranked listing of depth journals by number of appearances in combined random sample for 302 searches

Journal	Appearances	Journal	Appearances
Nature (London)	90	J Clin Invest	23
Brit Med J	84	Cancer	22
Ann NY Acad Sci	82	Deutsch Med Wschr	22
JAMA	74	Endocrinology	22
Lancet	74	Pediatrics	22
Proc Soc Exp Biol Med	61	Arch Gen Psychiat (Chicago)	21
New Eng J Med	54	Bibl Haemat	21
Amer J Obstet Gynec	49	Exp Cell Res	21
Amer J Physiol	49	J Bact	21
Biochim Biophys Acta	45	J Clin Endocr	21
Canad Med Ass J	44	J Exp Med	21
Ann Intern Med	35	Proc Nat Acad Sci USA	21
Arch Intern Med (Chicago)	35	Surgery	21
Arch Surg (Chicago)	34	Virology	21
Amer J Surg	33	Acta Chir Scand	20
Science	33	J Molec Biol	20
J Appl Physiol	31	J Thorac Cardiov Surg	20
Arch Ophthal (Chicago)	30	Presse Med	20
Proc Roy Soc Med	30	Radiat Res	20
Acta Un Int Cancr	29	Amer J Clin Path	19
Amer J Ophthal	29	Amer Heart J	19
Ann Surg	29	Amer J Psychiat	19
Circulation	29	Amer J Path	19
Amer J Dis Child	28	C R Acad Sci (D) (Paris)	19
Biochem J	28	J Immun	19
Radiology	28	J Pharmacol Exp Ther	19
Acta Endocr (Kobenhavn)	27	Surg Gynec Obstet	19
Amer J Roentgen	27	Circ Res	18
Dis Chest	27	J Cell Biol	18
J Neurosurg	27	Metabolism	18
Obstet Gynec	27	Neurology	18
J Biol Chem	26	Biochem Pharmacol	17
Amer J Cardiol	25	Experientia	17
Cancer Res	25	J Endocr	17
Fed Proc (Transl Supp)	25	J Nat Cancer Inst	17
J Pediat	25	J Obstet Gynaec Brit Comm	17
Acta Med Scand	24	J Nutr	17
Amer J Med	24	J Physiol (London)	17
J Lab Clin Med	24	J Urol	17
Med J Aust	24	S Afr Med J	17
Surg Forum	24	Amer Surg	16
Arch Derm (Chicago)	23	Biochem Biophys Res Commun	16
Arch Neurol (Chicago)	23	Geriatrics	16

Table 34

(continued)

Journal	Appearances
J Neurol Neurosurg Psychiat	16
Southern Med J	16
Acta Neurol Scand	15
Arch Path (Chicago)	15
Brit Heart J	15
Exp Neurol	15

Six journals with 14 appearances each

Six journals with 13 appearances each

Ten journals with 12 appearances each

Twelve journals with 11 appearances each

Sixteen journals with 10 appearances each

25 journals with 9 appearances each

21 journals with 8 appearances each

25 journals with 7 appearances each

37 journals with 6 appearances each

43 journals with 5 appearances each

59 journals with 4 appearances each

74 journals with 3 appearances each

110 journals with 2 appearances each

116 journals with 1 appearance each

Table 35

Ranked listing of non-depth journals by number of appearances in combined random sample for 302 searches

Journal	Appearances
Practitioner	19
Nederl T Geneesk	15
Prensa Med Argent	14
New York J Med	13
Orv Hetil	11
Biomed Sci Instrum	10
Deutsch Gesundh	10
J Ass Physicians India	10
Med Welt	10
Pediatriia	10
Pol Tyg Lek	10
J Indian Med Ass	9
Mod Hosp	9
Postgrad Med	9
Cancer Chemother Rep	8
Eksp Khir Anest	8
USAF 6570 Aerospace Med Res Lab	8
Appl Ther	7
Bull WHO	7
Cesk Pediat	7
Gruzlica	7
J Lancet	7
Med Times	7
Naika	7
Postgrad Med J	7
Progr Cardiov Dis	7
Sovet Med	7
Un Med Canada	7
US NASA	7

17 journals with 6 appearances each
32 journals with 5 appearances each
32 journals with 4 appearances each
80 journals with 3 appearances each
165 journals with 2 appearances each
380 journals with 1 appearance each

EFFECT OF MEDLARS RESPONSE TIME

All requesters were asked whether the MEDLARS response time was satisfactory or whether processing delays had significantly reduced the value of the MEDLARS search. Only 21 of the 302 searches (7.0%) were significantly reduced in value by processing delays. This is considerably fewer than expected on the basis of the pretest results, which indicated that as many as 15% of all MEDLARS searches might be significantly reduced in value by unsatisfactory response time.

However, the pretest was conducted in early 1966, when the MEDLARS throughput time was highly unsatisfactory. General improvements in searching efficiency have led to a substantial decrease in average response time for MEDLARS searches, as the figures in Table 36, based only on test searches processed at NLM, demonstrate.

Table 36

Mean response time* of 198 MEDLARS test searches processed at NLM, by month in which the request was received at NLM.

1966		1967	
August:	57 days	January:	15 days
September:	64 days	February:	23 days
October:	56 days	March:	14 days
November:	39 days	April:	14 days
December:	24 days	May:	24 days
		June:	13 days

* "Response time" is calculated from the day a request is received at NLM to the day the results are mailed to the requester.

THE SERENDIPITY VALUE OF MEDLARS SEARCHES

The Form for Document Evaluation was designed to gather some data on the serendipity value of MEDLARS searches. In the case of articles judged of no value in relation to the information need prompting a MEDLARS request, the requester was asked to indicate whether or not he was glad to learn of the existence of the article "because of some other need or project".

The serendipity value will, of course, vary from requester to requester. In search # 3, four of the five irrelevant items (80%) were said to be of interest in relation to some other project; in search # 68 the serendipity ratio was 13/16 (81%); in search # 6 it was 8/20 (40%); and in search # 11 it was zero (0/26).

It may be that a search with a high serendipity score indicates that the requester has a high recall need: he is anxious to obtain as much related material as possible and is glad to see even very peripherally related items. A low serendipity score, on the other hand, indicates a requester with a very precise need: he wants only articles directly bearing upon the subject of his request and is not particularly interested in browsing in related areas.

The combined random sample (see Tables 32 and 33) for the 302 searches included 2999 articles judged "of no value" in relation to the information needs prompting the MEDLARS requests. The requesters indicated that they were glad to learn of the existence of 532 (18%) of these articles in relation to some other need or project.

However, this is undoubtedly an inflated estimate of the serendipity value of the machine search. This figure is based on the requester's evaluation of actual articles supplied to him in the photocopy form. It is not unlikely that, under these conditions, he would find some items of interest although not directly of value in relation to his current requirements. It is unlikely that, on the basis of the MEDLARS citation printout only, the requester would be sufficiently interested in 18% of the "irrelevant" citations to go to the trouble of acquiring and making use of the articles.

OUTPUT SCREENING

An experiment was conducted to determine how closely relevance predictions made at NLM on the basis of full citations plus index terms (as contained in a demand search bibliography) would coincide with actual relevance assessments made by requesters on the basis of seeing the articles themselves. For ten of the test searches, an M. D. on the staff of NLM made relevance predictions on the citations selected from the search printout by the random sampling procedure. An experienced MEDLARS search analyst did the same thing for a different group of 19 searches. Both the M. D. and the search analyst worked from marked copies of the MEDLARS printouts for the 29 searches. Besides the printout, they were given only (a) a copy of the original request, and (b) a copy of the search formulation. Both were given the same instructions, namely to delete citations that appeared obviously irrelevant. They were to leave in any citations about which they were doubtful, on the grounds that in general it is better to send too much than too little. These relevance predictions were then compared with the real-life value judgements made by the requesters when seeing the actual articles. The results of this experiment are somewhat depressing. Over ten searches the M.D. screened out about 6.7% of the irrelevant material and would thus have improved the average precision for these searches from 49.6% to 56.3%. Unfortunately, his screening would also have eliminated 8.6% of the relevant material.

These results were almost exactly replicated by the search analyst, who would have improved precision from 48.3% to 53.4%, but at the same time would have eliminated 8.5% of the relevant material. In other words, the screening operation reduced rather than improved the search results, dropping the average recall by more than it raised the average precision.

This prompted the author to carry out an analysis, based on the 19 searches screened by the searcher, to determine why the screening operation was so unsuccessful. In some cases the screener followed the errors of the original searcher (e.g., in search # 177 both were incorrect in thinking that ABRUPTIO PLACENTAE was relevant to the topic of "premature rupture of the fetal membranes"), which prompts the conclusion that, if screening is done, it should be done by a second person and not by the original searcher. In this case, the screener did not spot the error, but another screener might have done. The original searcher would almost certainly not have noticed it.

However, the principal problem is the fact that titles and tracings are frequently inadequate as indicators of content. In fact, the tracings may confuse rather than assist the screening operation. In some cases it appears that the screener accepted or rejected an article on the basis of the terms assigned, while the opposite decision might have been taken on the title alone. The index terms do not indicate the extent to which a

topic is treated in an article; nor do they indicate (except where subheadings are used) relationships between various items of subject matter.

The fact that titles are not good content indicators would suggest that MEDLARS users may well overlook a substantial number of the potentially relevant articles cited in a search printout. For example, consider once more the much-quoted search (#177) on premature rupture of the fetal membranes. The requester judged eight articles as relevant on the basis of seeing complete photocopies. The titles of these eight articles are as follows:

1. Perinatal mortality and active labor conditions.

2. Incidence of maternal and fetal complications associated with rupture of the membranes before onset of labor.

3. Prognosis in premature rupture of the membranes.

4. Pneumococcal laryngitis in the newborn infant

5. Conservative treatment of threatened premature labor

6. Cervical flora in patients with premature rupture of membranes.

7. Current concepts on premature rupture of the fetal membranes.

8. Bacterial shock . . . after rupture of the membranes three days previously

The articles fall into three groups: the group comprising # 2, 3, 6, 7, and 8 which, from the titles, appear obviously relevant; article 5, which may or may not be relevant as judged by the title; and articles 1 and 4, the titles of which give little indication of their relevance to the subject of premature rupture. Since these last three articles did not have the term FETAL MEMBRANES assigned to them, it would be reasonable to suppose that the requester would have judged only 5/8 relevant on the basis of the search printout. The implications of this are obvious. MEDLARS may be retrieving an average of 58% of the relevant articles within its base, but the proportion of the relevant literature brought clearly to the attention of the requester (to the point that he would be likely to obtain a copy of the full article) could well be much less than this.

This finding led to a second experiment. In this, the search analyst who had previously screened 19 searches was given, some two months later, nine of these searches to screen once more. This time, however, she worked from titles and abstracts, but no tracings. Abstracts or summaries

were obtained from the articles themselves* or from an abstracting publication such as Biological Abstracts or Excerpta Medica.

The results for the nine searches are presented below:

	Actual search precision	Screened precision based on citations and tracings	Screened precision based on abstracts
# 34	13/27 = 48.1%	12/20 = 60.0%	8/11 = 72.7%
# 36	13/18 = 72.2%	13/17 = 76.5%	8/8 = 100%
# 59	16/22 = 72.7%	15/18 = 83.3%	14/14 = 100%
# 105	0/10 = 0	0/10 = 0	0/8 = 0
# 177	3/23 = 13.0%	3/22 = 13.6%	3/7 = 42.8%
# 215	14/24 = 58.3%	11/14 = 78.6%	11/14 = 78.6%
# 220	4/7 = 57.1%	3/3 = 100%	4/7 = 57.1%
# 240	6/20 = 30.0%	6/18 = 33.3%	5/12 = 41.7%
# 245	15/24 = 62.5%	15/24 = 62.5%	13/17 = 76.5%
AVERAGE	45.9%	56.4%	63.3%

This particular group of nine searches had a rather low average precision ratio of 45.9%, and the initial screening was able to improve the average precision to 56.4%. However, a proportion of the relevant literature was also screened out (for example 1/13 in search # 34, 1/16 in # 59, and 3/14 in # 215), the average loss over the nine searches being 6.7%. Using abstracts, the screener was much more drastic in deleting articles and was thus able to raise the precision ratio to 63.3%. However, more of the relevant literature was also eliminated (5/13 in # 34, 5/13 in # 36, 2/16 in # 59, for example), the average loss over the nine searches being 15.7%.

Unfortunately, the situation appears to be as follows:

1. Given only titles and tracings, the screener is cautious and does not eliminate very much. However, as we have shown, titles and tracings are frequently rather poor indicators of content, and the screener

---

* It is in itself interesting to note that, of the 175 articles involved, no less than 170 (97%) contained an abstract, summary or conclusions section that was a fair indicator of content (although not all, of course, were in English).

eliminates a significant amount of the relevant material while improving precision a few percent.

2. Given abstracts, the screener becomes more bold. Further irrelevant material is screened out. However, interpretation of the request becomes more important with the additional indication of content. If the screener misinterprets the requester's requirements, the very fact that abstracts are available will tend to cause more of the relevant material to be discarded. In fact, because a requester's value judgements are personal and because, as we know full well, requests rarely completely coincide with information needs, the screener would be unlikely to closely match the requester's assessments even on the basis of full texts.

INDEXER CONSISTENCY

As already reported, eighteen articles from non-depth journals, unretrieved in test searches, were re-indexed to determine whether more exhaustive indexing would have allowed their retrieval. The fact that this re-indexing was done, allows us to carry out a small analysis of indexer consistency. Only the sixteen articles for which we have three versions of the re-indexing were included in this analysis.

Each of the three re-indexings for the sixteen articles wasthe work of an indexer-reviser pair. That is, a comparatively inexperienced indexer assigned terms, but these were later checked by a senior indexer ("reviser"). Under these circumstances one would expect the consistency level to be much higher than the consistency level for indexing that is not revised. Because of the wide discrepancy, in the number of terms assigned, between the original non-depth indexing and the three versions of the re-indexing, the former was not included in the consistency study. For the purposes of this analysis, a main term/subheading pair was counted as a single term, and was differentiated from the main term on its own. Thus ADENINE NUCLEOTIDES/METABOLISM was accepted as a term different from the term ADENINE NUCLEOTIDES alone.

The measure of consistency adopted was that defined by Rodgers[3] and by Hooper[4]. The consistency of a pair (CP) of indexers (i.e., the consistency of one indexer with respect to a second), in the indexing of a particular article, is based on the number of index terms used in common by the two indexers, divided by the total number of terms used by either indexer.

That is,

$$C\,P\ (\%) = \frac{100\ A}{A + M + N}$$

where, A = the number of term agreements between "M" and "N" for a specific article.

M = the number of terms used by "M" but not used by "N".

N = the number of terms used by "N" but not used by "M".

Because we have three versions of the indexing, for each of the sixteen articles, we must compute three consistency pairs (CPs) for each article (A and B, A and C, B and C) and average these results to arrive at an overall consistency ratio for the three versions. We can then average these consistency values over all sixteen articles.

Consider a particular article, #3. Between indexing A and indexing B there are five distinct MeSH terms, three of which are common to the two

indexings, so that the consistency score is 3/5, or 60%. The consistency score for A and C is 33.3% and the consistency score for B and C is 50%. When we average these, we arrive at 47.8% as the average inter-indexer consistency ratio for the indexing of this article. The complete list of these ratios is as follows:

Article	#1	41.9%
"	#2	40.2%
"	#3	47.8%
"	#4	23.0%
"	#5	25.1%
"	#6	34.0%
"	#7	28.0%
"	#8	12.8%
"	#9	28.0%
"	#10	57.5%
"	#11	22.1%
"	#12	17.3%
"	#13	22.3%
"	#14	32.8%
"	#15	63.7%
"	#16	53.5%

The overall consistency average is 34.4%.

Although we have nothing with which we can legitimately compare it, a consistency ratio of 34.4% seems rather low. In the various tests reported by Hooper the consistency values vary from 10% to 80%. However, the low values reported were mostly achieved in tests involving "free" indexing (i.e., without a controlled vocabulary). Factors affecting consistency will include the degree of vocabulary control, the size and specificity of the controlled vocabulary, the average number of terms assigned in indexing, and the "hardness" or "softness" of the subject matter being indexed. One would have expected a fairly high consistency value in the present study because a controlled vocabulary is being used, and each indexing had the normalizing influence of a revision process.

The author looked closely at these samples of re-indexing to see if any factors contributing to low consistency could be isolated. Some of the variations were extraordinary. In the case of article #8, for example, indexer A and indexer B agreed on only one term out of twelve, and that term was the check-tag HUMAN! In article #4, A regarded ALBUMINURIA as highly important because this term was designated as a print term. Yet indexer B did not even include it among 18 terms assigned. Even check-tags, about which there should be little disagreement, were not assigned consistently. In article #4, indexer A assigned only the check-tag HUMAN: B assigned ADOLESCENCE, ADULT, AGED, CHILD, FEMALE, HUMAN, MALE and MIDDLE AGE: C assigned ADOLESCENCE, ADULT, CHILD, HUMAN and MIDDLE AGE.

Indexer A and indexer C regarded article #5 as being primarily related to physical therapy. Indexer B, however, considered it highly related to psychotherapy.

There are very few data processing terms in MeSH so one would expect a high level of consistency in this area. Yet, in indexing article #13, A used AUTOMATION, B used AUTOMATIC DATA PROCESSING, and C used both. B regards the tools used as COMPUTERS while A regards them as COMPUTERS, DIGITAL. Likewise in article #14: A used COMPUTERS, while B and C used COMPUTERS, DIGITAL. Indexer A also assigned the geographical heading MISSOURI to this article, but this was not used by B or C.

The highest consistency figures were achieved on articles #10 (57.5%), #16 (53.5%) and #15 (63.7%). There are reasons for this. Articles 10 and 16 involve areas somewhat peripheral to the main stream of MEDLARS, and for which there are comparatively few terms available in the vocabulary. Since there are only a few fairly general terms to choose from, indexers are likely to show greater consistency. Although article #15 was indexed more exhaustively than the others, the consistency ratio was highest of any (63.7%). This is because it deals with a "hard" area, involving specific chemicals and enzymes named in the article. Since specific terms are available for these substances, a high level of consistency is possible

It was noted in analysis, however, that the use of subheadings contributes substantially to the low consistency level. If we ignored the subheadings (by accepting, for example, MATERNAL-FETAL EXCHANGE as the same term as MATERNAL-FETAL EXCHANGE/PHYSIOLOGY, which indeed it is when we carry out a search without subheadings), the consistency ratio would improve somewhat. In fact, when we recalculate the consistency ratio, ignoring subheadings, we arrive at the average consistency ratio of 46.1% for the sixteen searches. From this we must conclude that, although subheadings add greatly to the specificity of the vocabulary, and have great potential in reducing both false coordinations and incorrect term relationships, like role indicators[5] they are difficult to apply consistently.

REQUESTS REJECTED BY MEDLARS

During the period of the test, a record was made of any request submitted to NLM by mail from the cooperating organizations, that were rejected by MEDLARS. There were seventeen in all (seventeen out of approximately 270 requests received by NLM from these organizations, in the test period, indicates a rejection rate in the region of 6%) and these divide up naturally into three groups:

1. Eight were rejected on the grounds that they did not "necessitate the facilities of the computer to sort or correlate a complex series of variables". These are single-faceted requests, involving single MeSH terms, and the Search Section felt that a machine search could add little to a conventional search in Index Medicus. While it is legitimate to reject a request on these grounds, there appear to be no consistent standards governing acceptance or rejection. For example, a search on "pathogenesis, diagnosis and treatment of hydrocephalus" was rejected, while another on "experimental production and known causes of hydrocephalus" was accepted. Since the latter search was conducted on the single term HYDROCEPHALUS, it is difficult to understand in what way it differs from the rejected search. Likewise, a request on "toluidine blue dye" was rejected while another on the Bodian protargol method of silver staining (searched on the single term SILVER PROTEINS) was accepted.

2. Eight were rejected because no specific MeSH terms exist to express the subject matter of the request, and there is nothing that can be done by term coordination to get at the specific topic of interest. Such requests, insofar as they are regarded as falling within the scope of the system, indicate vocabulary inadequacies, and should routinely be input to the MeSH group at NLM.

3. One search was rejected as being outside the scope of the system.

PART 3

CONCLUSIONS AND RECOMMENDATIONS

The MEDLARS evaluation, discussed in this report, was a complete system evaluation inasmuch as it studied all components affecting the performance of the system as measured by the satisfaction of MEDLARS users. The benefit of this type of evaluation program lies not in detecting specific failures, but in identifying kinds of failures that are prone to occur, and indicating in which areas corrective action is most urgently needed.

Overall MEDLARS performance

The test results have shown that the system is operating, on the average, at about 58% recall and 50% precision. On the average, it retrieves about 65% of the major value literature in its base at 50% precision. However, as previously noted, averages are somewhat misleading in this context. Few of the individual search results fall in the area bounded by the average ratios ± 5%. In fact, the results are widely scattered. Some of the searches appear to have performed very well, with high recall accompanied by high precision. Other searches achieved completely unsatisfactory recall results. The most important factors governing the success or failure of a MEDLARS search were discussed in some detail in Part 2 of this report.

The MEDLARS average performance ratios may seem low when compared with certain figures (e.g., 90% recall at 90% precision) quoted in the documentation literature. Unfortunately, the great majority of the quoted figures are completely without foundation. There is no other fully operational retrieval system, of any significant size, that has exposed itself to the rigours of an evaluation program such as the one here reported. The author considers it extremely unlikely that any other large mechanized retrieval system, if it were evaluated in the way that MEDLARS has been evaluated, would be found to be operating at a higher average performance level.

It should be borne in mind, in considering the MEDLARS figures, that the present evaluation has been conducted as stringently as possible. The author has assumed the role of an impartial (but hopefully constructive) critic of MEDLARS. Whenever a decision had to be made, it was made against the system. An article judged "of value" by the requester was accepted as being "relevant" even though it was found to contain very slight reference to the subject of the request. Known relevant articles that were not retrieved were counted against the system, even in cases in which the requester, in agreeing to the exclusion of certain terms, was himself largely responsible for the misses.

It must also be remembered that "relevant", within the context of this program, has been defined as "of value to the requester in relation to the information need prompting his request to the system". Relevance to an information need is very different from relevance to a stated request. In fact, had we evaluated MEDLARS on the basis of the latter criterion, both recall ratio and precision ratio would have been approximately 10%

higher, because we would not have counted against the system the 25% of the recall failures and 17% of the precision failures presently attributed to inadequate user-system interaction.

To counterbalance the stringency of the evaluation, we have to recognize the fact that the analysts preparing search formulations for the various test requests were aware that these searches were subsequently to be evaluated. Almost certainly there was some "spotlight" effect. We can therefore say that the present evaluation has studied the performance of MEDLARS with one component of the system (namely search formulation) behaving optimally. There could also have been some "spotlighting" in the area of user-system interaction. However, as we know, this might have degraded performance rather than improved it.

Figure 7 and Figure 15 present performance curves for the MEDLARS test searches, the former based on performance points for the various centers, the latter on performance points for the 6-5-4 subsets in 118 searches. By extrapolation, we can hypothesize a generalized MEDLARS performance curve looking something like that of Figure 17. From results of other investigations, largely on experimental or prototype systems, using Cranfield-type methodology, we expected (before the study was conducted) that MEDLARS would be performing rather differently than it was actually found to be. In fact, the author expected that the system would function in a high recall, low precision mode in the region, say, of 75-90% recall at 10-20% precision. The results actually achieved over 300 test searches do not indicate a performance worse than expected, but they do indicate a performance different from that expected.

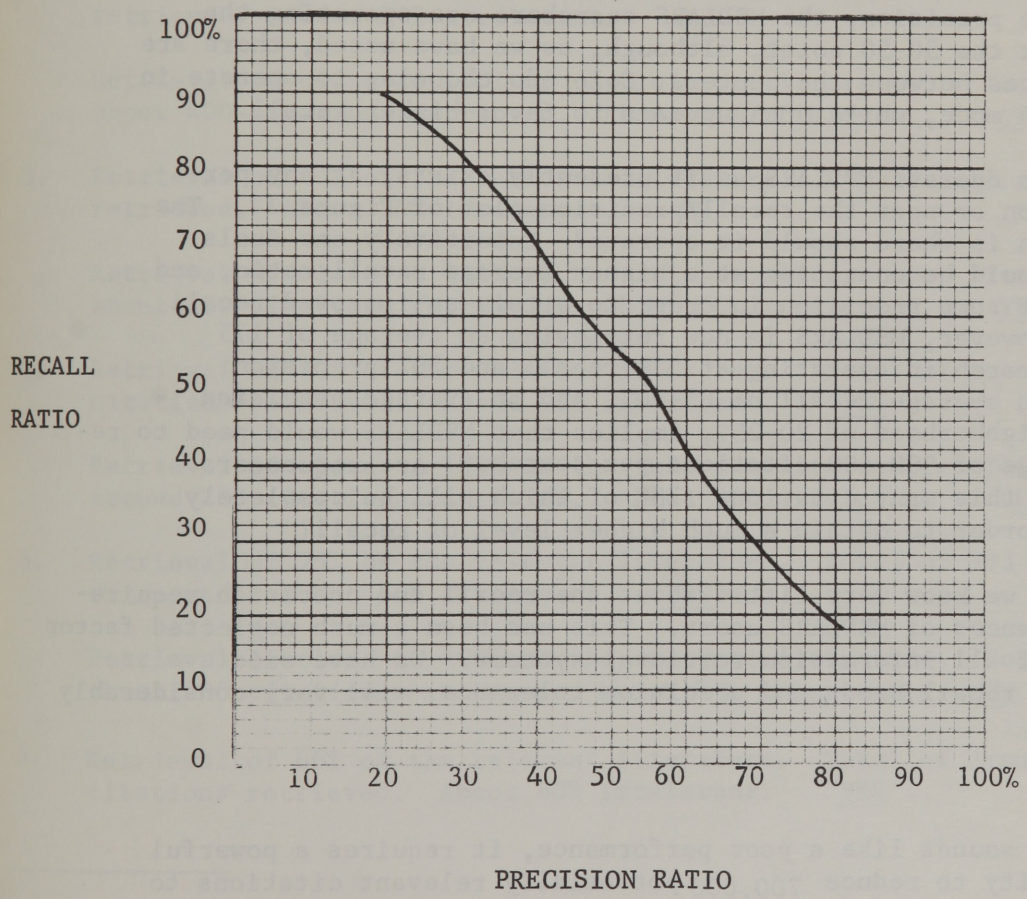
The fact that, on the average, MEDLARS is operating at 58% recall and 50% precision, indicates that, consciously or unconsciously, the MEDLARS searchers choose to operate in this general area. It would be possible for MEDLARS to operate at a different performance point on the recall/precision curve of Figure 17. By broadening of search strategies one could obtain a much higher average recall ratio, but this could only be obtained at a lower average precision ratio. However, the indications are that MEDLARS could operate in a high recall mode (say 80-90%) at a much higher precision ratio than we could have expected on the basis of other evaluations conducted by means of Cranfield-type methodology.

Obviously, it is always possible to achieve 100% recall for any request by retrieving the entire data base. This is nonsense in that, under these conditions, the filtering capacity of the system is not being brought into play at all. With sufficient broadening of each search strategy, however, it would be possible for MEDLARS to achieve very close to 100% recall for any request without retrieving the entire collection. However, in some searches 100% recall (or close to 100%) can be achieved at a tolerable precision ratio, while in other searches we cannot approach 100% recall and still obtain acceptable precision.

Consider once more the search (# 194) on nutritional aspects of chromium, and the search (# 177) on premature rupture of the fetal membranes. If we conducted the former search on the single-term strategy CHROMIUM or CHROMATES we would obtain 95% recall (we fall short of 100% because of indexer omissions) and retrieve in the neighborhood of 100 documents

Figure 17

Generalized MEDLARS performance curve



which about one third are relevant. In this case, we can assure the requester of almost maximum recall and still operate at a tolerable precision ratio. It does not seem too unreasonable to expect the requester to examine 180 citations in order to find 50-60 of some value to him.

On the other hand, because of indexing omissions and inadequacies of the index language, we could only approach 100% recall in the search on rupture of the membranes by searching on FETAL MEMBRANES and also on all terms relating to pregnancy complications, labor complications, and newborn infant disease. This would retrieve several thousand citations of which only about 30-35 would be in any way relevant. Almost certainly we could not expect the requester to examine several thousand citations in order to find 30-35 pertinent ones (especially since we know, from the analysis of output screening, that the requester in unlikely to be able to recognize all the relevant items anyway.)

The conduct of a machine search is essentially a compromise between recall and precision. In attempting to obtain a satisfactory recall at an acceptable precision, the MEDLARS searchers are operating the system almost at the 50-50 point, although, as we have noted, there are policy differences between the centers, Colorado choosing to operate in a high precision mode, while UCLA appears to favor higher recall.

We can choose to operate MEDLARS, as it presently exists, at any performance point on or near the recall/precision plot of Figure 17. The crucial question is where should it operate? Intuitively one feels that MEDLARS should be operating at a higher average recall ratio, and should sacrifice some precision in order to attain an improved recall performance. However, MEDLARS is now retrieving an average of 175 citations per search in operating at 58% recall and 50% precision. To operate at an average recall of 85-90%, and an average precision ratio in the neighborhood of 20-25%, implies that MEDLARS would need to retrieve an average of 500-600 citations per search.* Are requesters willing to scan this many citations (75% of which will be completely irrelevant) in order to obtain a much higher level of recall?

In actual fact, we know very little about the recall and precision requirements and tolerances of MEDLARS users. This has been a much neglected factor in the design of all information retrieval systems. We have said previously that recall needs, and precision tolerance, will vary considerably

---

* Although this sounds like a poor performance, it requires a powerful filtering capacity to reduce 700,000 potentially relevant citations to 600 potentially relevant, without losing a significant amount of the relevant literature.

from requester to requester, depending upon the purpose of the search. Out of curiosity, the author wrote to ten scientists, participating in the evaluation, with a view to determining their actual recall needs and precision tolerances. In each case, through search analysis, we knew roughly how each search had performed and had also made some hypotheses on how many citations would need to be retrieved in order to approach 100% recall. In each case, the requester was asked to indicate whether he was satisfied with the level of performance achieved or whether he would have tolerated a much lower precision in order to get somewhere near to 100% recall. A specimen letter is included as Figure 18, and the answers of the eight respondents are tabulated below:

1. Retrieval of 33% of the relevant literature. Total of 25 citations retrieved. About 30% irrelevant.* YES

   Retrieval of close to 100% of the relevant literature. Total of about 100 citations retrieved. About 75% irrelevant. NO

2. Retrieval of 77% of the relevant literature. Total of 233 citations retrieved. About 80% irrelevant. NO

   Retrieval of close to 100% of the relevant literature. Total of about 400 citations retrieved. About 90% irrelevant. YES

3. Retrieval of 40% of the relevant literature. Total of 15 citations retrieved. About 10% irrelevant. NO

   Retrieval of close to 100% of the relevant literature. Total of about 100 citations retrieved. About 50% irrelevant. YES

4. Retrieval of 60% of the relevant literature. Total of around 100 citations retrieved. About 95% irrelevant. YES

   Retrieval of close to 100% of the relevant literature. Total of around 250 citations retrieved. About 95% irrelevant. NO

5. Retrieval of 75% of the relevant literature. Total of 333 citations retrieved. About 40% irrelevant. YES

   Retrieval of close to 100% of the relevant literature. Total of about 500 citations retrieved. About 50% irrelevant. NO

6. Retrieval of 66% of the relevant literature. Total of around 400 citations retrieved. About 60% irrelevant. YES

---

* In each case, the first alternative posed represents the performance actually estimated for the system.

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
8600 WISCONSIN AVENUE
BETHESDA, MD. 20014

NATIONAL LIBRARY OF MEDICINE

REFER TO: NLM - R & D

October 30, 1967

Department of Anesthesia
U. S. Naval Hospital
National Naval Medical Center
Bethesda, Maryland 20014

Dear

You will remember that recently we conducted a search for you on the subject of repair of amputated finger tips, and that you very kindly assisted us in evaluating the results of this search. There is one more thing that you could help us with if you would be so good.

We need to know something of the requirements and tolerances of MEDLARS users. From my evaluation, I believe that the MEDLARS search retrieved only about 33% of the relevant articles on the precise topic of your interest. However, to retrieve anything approaching 100% of the relevant literature I believe that we would have needed to retrieve many more citations in total - possibly about 100, of which only about 25% would be directly relevant.

The question is: Would you have preferred to look through the additional irrelevant citations in order to approach 100% retrieval of the relevant literature?

If you could please return this letter, marked with your answer, I should indeed be most grateful.

Would prefer (delete whichever inapplicable):

1. Retrieval of 33% of the relevant literature. Total of 25 citations retrieved. About 30% irrelevant.

2. ~~Retrieval of close to 100% of the relevant literature. Total of about 100 citations retrieved. About 75% irrelevant.~~

Sincerely,

F. Wilfrid Lancaster
Information Systems Specialist
Research and Development
National Library of Medicine

Figure 21

6. Continued.

Retrieval of close to 100% of the relevant literature. Total of at least 700 citations retrieved. About 70% irrelevant. NO

7. Retrieval of 66% of the relevant literature. Total of 190 citations retrieved. About 50% irrelevant. YES

Retrieval of close to 100% of the relevant literature. Total of about 300 citations retrieved. About 60% irrelevant. NO

8. Retrieval of 36% of the relevant literature. Total of 10 citations retrieved. About 60% irrelevant. NO

Retrieval of close to 100% of the relevant literature. Total of about 60 citations retrieved. About 80% irrelevant. YES

One cannot draw firm conclusions on the basis of eight responses of this kind. Nevertheless, the results are very interesting. It appears that we are wrong in assuming that most requesters want maximum recall. Five of these eight respondents have indicated satisfaction with the less-than-maximum results. At least, they indicate unwillingness to examine additional irrelevant citations in order to approach 100% recall. In relation to these responses, the general performance level at which MEDLARS has chosen to operate would appear to be a reasonable compromise between recall and precision. However, no clear picture emerges from the responses. In # 1 the requester is satisfied with 33% recall and would not care to examine 100 citations, at 25% precision, in order to substantially improve on this recall figure. On the other hand, in # 2 the requester is prepared to examine 400 citations, 90% of which are irrelevant, in order to approach 100% recall.

Clearly, each individual has his own requirements in relation to the tradeoff between recall and precision, and we cannot generalize on this subject. It is important, therefore, that the MEDLARS demand search request form be so designed that it establishes for each request the recall requirements and precision tolerances of the requester, thus allowing the searcher to prepare a strategy geared as required to high recall, high precision, or some compromise point in between. The search request form will be mentioned again later.

Upgrading the performance of MEDLARS

So far we have considered how MEDLARS is operating. We have also indicated that the present system could choose to operate at some different average performance point on the recall/precision plot of Figure 17. However, this evaluation program has not been conducted primarily to determine the present performance level. Rather, it was conducted to discover what needs to be done to upgrade the performance of the present system,

i.e., what can be done to move the generalized performance curve of Figure 17 further to the right in order to achieve a higher average performance capability (e.g., 58% recall at 70% precision, 80% recall at 50% precision, 90% recall at 40% precision). The remainder of this report will be concerned with conclusions and recommendations relating to the various components of the MEDLARS demand search system.

In considering these recommendations, it must be recognized that, although we can do certain things to a system ostensibly to improve recall (e.g., indexing more exhaustively) and other things ostensibly to improve precision (e.g., increasing the specificity of the index language or introducing relational indicators), the present study has shown that there is no clear cut distinction between improving recall capabilities and improving precision capabilities. Recall and precision are strongly interconnected in an inverse relationship, and searching involves a compromise between the two. Therefore, inadequate precision devices can affect recall just as much as they affect precision. As an example, consider search # 93, relating to hypophosphatasia. HYPOPHOSPHATASIA is a fairly recent provisional heading, so the search had to be conducted at a more general level for the earlier material. Too avoid an unacceptably low precision ratio, the searcher was cautious in the formulation, using only

METABOLISM, INBORN ERRORS	and	BLOOD ALKALINE PHOSPHATASE ALKALINE PHOSPHATASE.

This strategy retrieved only six citations, all relevant, but we estimate that this is but a very small fraction of the total relevant literature. It would be necessary to generalize much more to BLOOD ALKALINE PHOSPHATASE alone (with over 800 postings) in order to obtain high recall. This, then, is clearly a situation in which lack of specificity in the vocabulary has led to recall failures rather than precision failures, and we can expect that recall would have reached an acceptable level had the specific term HYPOPHOSPHATASIA always been available.

Similarly, in search # 181 it was not possible to express asymptomatic proteinurias because no specific term exists for this notion. The searcher attempted to keep irrelevancy within bounds by negating kidney disease terms. Unfortunately, this screened out some of the relevant items also, and achieved 60% recall and 17.4% precision. Again, we would expect both better recall and better precision if an appropriate specific term were available in the vocabulary.

Like situations result from other compromise strategies designed to avoid false coordinations and incorrect term relationships, and we can thus safely say that, in the long run, a system change that adds greater precision capabilities will also tend to allow improved recall performance.

Regarding these conclusions and recommendations, the author has considered it his function to expose system weaknesses and point to work that needs

to be done and decisions that need to be taken. He has not considered it his present responsibility to carry these recommendations to the point of, for example, designing finished forms or proposing new specific subject headings.

It must also be borne in mind that changes made in the area of searching, or the area of user-system interaction can have immediate effect on the system . On the other hand, it will be some years before changes in indexing and index language can have a substantial effect on the complete data base.

User-system interaction

The greatest potential for improvement in MEDLARS exists at the interface between user and system. A significant improvement in the statement of requests can raise both the recall and the precision performance of the system: 25% of the MEDLARS recall failures and 16.6% of the precision failures are attributed, at least in part, to defective interaction.

We recommend that the search request form be completely redesigned along the lines proposed in Figure 11. It is obviously crucial to the success of a MEDLARS search that a request should accurately reflect the actual information need of the requester. For this reason, it is worth investing a substantial amount of time and effort in the design of a new request form. In particular, the questionnaires relating to search limitations and to the recall/precision tradeoff (parts 5 and 6 of the proposed form) will require very careful presentation and wording. The search request form will require testing in draft (possibly several drafts) before it is finally accepted and put into use.

We recommend that all requesters be required to complete this form personally, even in situations in which the requester makes a personal visit to a MEDLARS center or to his local library. In personal confrontation between requester and search analyst, the function of the latter should be to clarify the request statement, where necessary, but not to influence it. In particular, a request should not be discussed with a requester in terms of Medical Subject Headings, or at least not until the requester's own statement of need has been captured on the search request form.

The MEDLARS index language

We recommend a thorough re-appraisal of methods presently used to update Medical Subject Headings. In particular, we feel that the future success of the retrospective search function demands a shift in emphasis away from the external advisory committee on terminology and towards the continued analysis of the terminological requirements of MEDLARS users as reflected in the demands placed upon the system. As part of quality control procedures, the MeSH group, in cooperation with the Search Section, should undertake the continuous analysis of MEDLARS search requests with a view to identifying areas of weakness in MeSH and legitimate requirements that cannot presently be satisfied because of inadequate terminology.

We recommend that the MEDLARS entry vocabulary be regarded as an integral part of the index language of the system of no less importance than MeSH itself. The entry vocabulary, which should be the joint responsibility of the MeSH group and the Index Section, will require considerable improvement if it is to function adequately. Any significant topic, encountered by an indexer, for which there exists no specific MeSH term, is a candidate for inclusion in the entry vocabulary. However, we cannot expect NLM indexers, who are required to adhere to a tight production quota, to maintain an adequate entry vocabulary. It should be the function of the indexers to "flag" topics that require a new MeSH term, provisional heading, or entry vocabulary term, for subsequent analysis and action in the MeSH group.

The present format of the entry vocabulary, as it exists in the shape of an Authority File on 3 x 5" cards, should be replaced by an alternative amenable to (1) machine manipulation and updating and (2) rapid accessing by indexers and searchers. Every indexer and every searcher, including those at the centers, should be able to consult the entry vocabulary as easily as they can consult MeSH itself. This implies, at the present time, an entry vocabulary in book form. Consultation of a continuously updated entry vocabulary in an on-line browsing mode should be within the capabilities of the next generation system.

The introduction of subheadings, in 1966, appears to have been a most valuable improvement to the retrospective search function of MEDLARS as well as to the printed bibliographies. Subheadings afford an economical way of greatly increasing the specificity of the vocabulary. The use of subheadings can obviate the vast majority of the precision failures presently attributed to false coordinations and incorrect term relationships. However, subheadings, in allowing much greater specificity and the expression of complex relationships between terms, present problems in consistency of application. It is important that all subheadings be carefully defined, and that strict rules govern the conditions of their use. One great advantage of subheadings is that the searcher has the option of using them or not using them as the recall and precision requirements of a particular search dictate.

We recommend an expansion in the use of subheadings within MEDLARS, and support the present trend away from pre-coordinated terms (e.g., BLOOD PRESERVATION, LUNG TRANSPLANTATION) in Medical Subject Headings to the more flexible approach of optional pre-coordination, at the time of indexing, by means of subheadings. There is need for additional subheadings in the system. In fact, any fairly general notion, applicable to a large number of MeSH terms, is a good candidate for use as a subheading (e.g., PRESERVATION, which is potentially applicable to all tissue terms, and such terms as ACUTE and CHRONIC, which are potentially applicable to most disease terms). The author has not considered it his function to produce a list of new subheadings, although in Part 2 of this report he did recommend certain types (e.g., those relating to various characteristics of pathological conditions) that search analysis showed to be of great potential value to the system.

It is the joint responsibility of searchers and the Medical Subject

Headings group to determine what new subheadings could usefully be incorporated into the system. This can only be done, as it has been in this evaluation, by careful analysis of the types of requests put to the system (their specificity and the conceptual relationships involved), and of search failures occurring through lack of specificity, false coordinations, and incorrect term relationships. We see no need for the introduction of additional syntactical devices (e.g., links and roles) into the MEDLARS index language.

Finally, the search analyses have revealed the need for improved check-tags to describe types of articles. In particular, it is necessary that, in searching, we should have a simple and foolproof way of distinguishing experimental articles from clinical articles. We should also be able to distinguish single case studies from "large case series." Some requesters are willing to accept the latter, but not the former. Similarly, it would be very useful if articles could be identified by level of treatment: we should avoid supplying the researcher on a particular topic with a large number of fairly superficial articles written for the general practitioner.

The MEDLARS searching strategies

A significant number of recall failures have been attributed to the searcher failing to exhaust all reasonable approaches to retrieval. In the next generation system, careful consideration should be given to additional term displays that can be generated to assist the searching function. These displays would differ from the present tree structures in cutting across conventional genus-species hierarchies. They would resemble the ad hoc agglomerations of terms ("hedges") that at present tend to be collected by individual searchers for their own personal use. These are really pre-established searching strategies. They are most useful in covering "aspects" or "points of view" in relation to a main search topic (e.g., "nutritional aspects", "genetic aspects", "epidemiology"). Although such pre-established strategies may not be 100% transferable from search to search, they should nevertheless have fairly general applicability. For example, the terms coordinated with SPINA BIFIDA to express epidemiology of this anomaly should surely be the same as the terms coordinated with MONGOLISM to express epidemiology of this syndrome. Once agreement has been reached on a pre-established strategy for a particular generally-applicable concept, this strategy can be stored in machinable form and merely referred to, in a search formulation, by unique identifying number (in the same way that one can presently request an explosion on a particular tree structure). The repeated reconstruction, and copying down, of strategies for notions that tend to recur frequently in MEDLARS searches is considered to be most uneconomical.

The author is concerned about the increasing complexity of searching within MEDLARS. Each additional vocabulary change makes the searcher's task more difficult. In the design and planning of the next-generation system, it is recommended that a study be conducted on the feasibility of "automatic term replacement" to compensate for vocabulary changes. For

example, HALLERVORDEN-SPATZ SYNDROME became a provisional heading on 2/13/65. It is necessary to search on various other term combinations (e.g., BRAIN DISEASES and GLOBUS PALLIDUS and SUBSTANTIA NIGRA) to retrieve the earlier material. However, searchers should not be repeatedly burdened with the task of determining what term combinations have to be used to retrieve articles predating the specific term. This should be done, once and for all, at the time the new term is introduced into the vocabulary. Thereafter, the searcher should need to use only the most recent, specific term in the search formulation. A computer program should be written to automatically add the terms or term combinations, with appropriate date restrictions, necessary to retrieve the earlier material.

Vocabulary changes add to the complexity of searching, but some of the complexity appears to be self-inflicted. We have already demonstrated that wide variations in complexity of strategies exist between the various MEDLARS centers. It is difficult to generalize on this point, but, on strictly economic grounds, a simple-minded approach to searching is recommended in cases in which high recall can be obtained with a tolerable precision ratio. For example, the search on toxicity and nutritional aspects of chromium (# 194), if conducted on the single terms CHROMIUM or CHROMATES, could have achieved close to 100% recall (at least 95%), at a tolerable precision ratio of at least 33%, while retrieving only about 180 citations. It seems uneconomical to coordinate several hundred terms with CHROMIUM or CHROMATES, in an attempt to cover only the aspects mentioned in the request, and thereby achieve 60% precision in a total retrieval of 94. Presumably, the more complex the search formulation the more time it takes to prepare and the more likely it is to contain logical errors or inappropriate term combinations.

A searcher has the capability, by varying the specificity and/or exhaustivity of the formulation, to construct a strategy designed to achieve high recall (that we would expect to be accompanied by low precision) or one which is more a compromise strategy, sacrificing some recall to an improved precision ratio. At the present time the individual searcher makes a fairly arbitrary decision as to what type of strategy to adopt. Consequently, much time may be spent in constructing a comprehensive strategy in cases in which the requester would be satisfied with much less than 100% recall. If, as suggested, we can use the search request form to capture the recall/precision requirements and tolerances of users, the searcher should in future be able to prepare a formulation matched to these requirements and tolerances.

A substantial number of precision failures were attributed to lack of specificity in searching. It is recognized, however, that search generalization is often necessary in order to obtain satisfactory recall in a search. In a special analysis, we examined this question of search generalization: when it is justified, when not justified, and how it may best be accomplished. We also examined the use of weighted searching (on Index Medicus terms) as a useful means of compromising between recall

and precision. The results of these analyses, which give general pointers rather than standard rules, are presented in Part 2 of this report.

It has been shown that a search analyst, working from a citation printout, cannot make relevance predictions that will closely replicate the value judgements of the requester himself on seeing the actual articles. Consequently, we suggest that the detailed citation-by-citation examination of a search printout, by a search analyst, is not a particularly valuable expenditure of effort. It would seem more worthwhile to have each search (including printout, formulation and request statement) examined more generally by a second searcher with a view to identifying the gross errors that can occur (e.g., use of inappropriate term or term combination, the missing of a complete aspect of the request, or the use of faulty search logic).

The amount of search reformulation (approximately 24% in the present evaluation) that appears to take place at NLM is surprising. Presumably much of this reformulation is done after having seen the search printout. Yet we know that relevance predictions do not closely coincide with the value judgements of requesters. This casts serious doubt on the need for, and value of, such a high level of reformulation. We know of at least one search (# 44) in which the reformulation substantially degraded performance: it retrieved none of the nine known relevant articles, whereas the original would have retrieved 7/9. In other cases (e.g., search # 302, which was eventually conducted on the single term SYRINGOMYELIA), it is hard to understand why a straightforward search would require a second attempt at formulation, with an attendant delay of two months for the requester.

The most legitimate reason for reformulation would be a search spoiled by logical error or by the accidental use of an inappropriate term or term combination. More effort should be made to identify this type of error, which is an offspring of complex formulations, at an earlier stage in the searching process. A reformulation rate of 24% must represent a substantial investment in search analyst time.

Somewhat related to the matter of reformulation is the use of the 500 printout "ceiling" at NLM. As previously discussed, if a search is cut off after printing 500 citations (as it was in the case of 13 of the test searches), this indicates either (a) a substantial volume of literature on the subject of the request, in which case the requester may have legitimate need for a complete printout, or (b) a poor search formulation, in which case there may be a legitimate need to reformulate. We recommend a reappraisal of NLM policy with regard to both reformulation and the use of the search cutoff.

The MEDLARS indexing

The most difficult problem relating to indexing policy, in any system, is the decision as to what level of exhaustivity to adopt. That is, how many

terms, on the average, should we assign to a document? In Part 2 we presented many data relating to this question. These now require pulling together in an effort to arrive at some conclusions.

Approximately 20% of the MEDLARS recall failures are attributed to indexing that is insufficiently exhaustive, whereas only 11.5% of the precision failures were attributed to exhaustive indexing. On the surface, then, one would recommend increasing the exhaustivity of the indexing, to improve the recall potential of the system, rather than reducing exhaustivity. It is better to err on the side of additional terms. Without a fairly high level of exhaustivity, it is impossible to achieve a high average recall performance at a tolerable precision level. On the other hand, we can usually improve the precision of a search by employing more specific and/or exhaustive search formulations.

However, from the re-indexing experiments reported in Part 2, we have reason to suppose that:

1. Only a very much higher level of exhaustivity of indexing would allow the retrieval of a significant number of the relevant "depth" articles that are missed because they are not indexed with sufficient terms. Thirteen of these articles (originally indexed at an average of 7.2 terms) were re-indexed (at an average of 9.1 terms), but only two (15.4%) would have been retrieved on the re-indexing. In the other articles, the "relevant" section is very minor and would probably only be covered if the average term assignment was raised dramatically (say to 25-30 terms).

2. On the other hand, approximately 30-40% of all the relevant "non-depth" articles that are presently missed by MEDLARS searches would be likely to be retrieved if these articles were indexed with an average number of terms comparable to the "depth" average.

We also have reason to believe that, all other things being equal, the MEDLARS recall ratio for depth articles is 70% whereas the recall ratio is only 54% for non-depth.

Moreover, as previously noted in Part 2 of this report:

1. The division by journal into "depth" and "non-depth" creates indexing anomalies. Some of the "non-depth" articles are clearly under-indexed while some of the "depth" articles are clearly over-indexed.

2. Because of term limitations, some of the non-depth articles are indexed in such general terms that it is difficult to visualize a single search in which they would be retrieved and judged of value. In other words, these citations are merely occupying space on the citation file.

To recapitulate, we can say: a substantial number of recall failures occur due to lack of exhaustivity of indexing; a marginal increase in the average number of terms assigned to "depth" articles is unlikely to result in any significant recall improvement while a major increase is unjustified on economic grounds; raising the present "non-depth" level to the present "depth" level is likely to result in a 30-40% improvement in retrieval of relevant articles from non-depth journals; the present division of journals into "depth" and "non-depth" has led to indexing anomalies and to the situation in which non-depth articles occupy 45% of the file but account for only 25% of the retrievals; some of the non-depth articles are never likely to be retrieved and judged of value because they are indexed much too generally.

On the basis of the above, we recommend that the present distinction between "depth" journals and "non-depth" journals be abandoned. This does not mean that all articles from the present non-depth journals should be assigned an average of ten index terms. Rather, it means that each article should be treated on its own merit and sufficient terms should be assigned to index the extension and intension of its content. We see no justification for an overall increase in indexing exhaustivity at the present time.

Although few indexing errors (in the sense of incorrect term assignment) were discovered in the evaluation, a significant number of indexer omissions were encountered. Indexer omissions accounted for approximately 10% of all the recall failures. However, some of these indexer omissions appear to be largely due to lack of specific terms in the vocabulary. If no specific term is available for a concept, either in MeSH or in the entry vocabulary, an indexer is quite likely to omit it entirely (rather than trying to cover the topic in a more general way). We believe that indexer omissions will be substantially reduced as the entry vocabulary is improved.

Moreover, a very small spot-check (reported earlier) suggests that perhaps 25% of the failures attributed to indexer omission might not be the fault of the indexers, but might be due to the deletion of a term after the indexer has assigned it. This is further discussed below.

Computer processing

Computer processing was not a major culprit in causing retrieval failures in this study. However, one situation remains to be explained. As described in Part 2, it was possible to check back to the indexer data forms and flexowriter hard copy for four 1966 articles that were unretrieved, although relevant to various test requests, because of "indexer omissions". In the case of three of these articles, examination of the data form confirmed that an important term had not in fact been used by the indexer. However, in the fourth case, the term which the indexer had been accused of omitting (PARATHYROID GLANDS) did in fact appear on the data sheet; it also appeared on the flexowriter hard copy. The term was used twice in indexing, once with the subheading DRUG EFFECTS and once

with the subheading CYTOLOGY. This citation was printed in the December 1966 Index Medicus, and again in the Cumulated Index Medicus, under both main heading/subheading combinations. However, a computer printout of the tracings for this citation now reveals that the term PARATHYROID GLANDS has since been completely deleted.

This deletion probably occurred during some file maintenance procedure. The important question is how did it occur and, more importantly, how often does inadvertent term deletion take place during file maintenance procedures? Unfortunately we have no idea of the possible magnitude of this problem at the present time. This could be the only citation in which this inadvertent deletion has occurred. On the other hand, it could be one of 1000 or even 10,000 cases. We recommend that a separate investigation be made to determine the effect of file maintenance procedures on file integrity in order that the cause and magnitude of this problem can be determined.

The relationship between indexing, searching and MeSH

The tendency towards compartmentalization of indexing, searching and MeSH has been noted. This is evident in the following: request analysis and search failure analysis have not been major inputs to MEDLARS vocabulary control; the entry vocabulary, which should be an integral part of the MEDLARS index language, and an essential tool of both indexers and searchers, has been neglected; searchers are not completely aware of indexing policies and conventions; the average indexer has little idea, as far as the demand search function is concerned, of what he is indexing for (i.e., the types of requests that are made of the system).

We recommend that the Library take steps necessary to achieve a close integration between the functions of indexing, searching and vocabulary control. (The writer has not considered it within his present frame of reference to recommend specific organizational changes, nor to study methods whereby such integration can be effected most efficiently and economically.) Although consistency problems may result at first, the present trend towards combining, at MEDLARS centers, the indexing and the searching functions, is considered to be a valuable move in the right direction.

Use of foreign language material in MEDLARS

The comparatively small use made of foreign language material, by demand search requesters, was observed in Part 2. While foreign language articles consume approximately 45% of MEDLARS input costs, we estimate that they contribute no more than 16% of the total demand search usage (i.e., no more than 16% of the articles retrieved and judged of value are in languages other than English).

It is difficult to make specific recommendations on this subject, apart

from urging that NLM re-evaluate in general its policies relating to foreign literature. Many requesters complained that translation services are not available to them or that translation is too costly. If NLM continues to devote 45% of its input effort to the foreign material, it might consider adopting a more active role in the translations area (perhaps by acting as a clearinghouse for translations in biomedicine).

The search printout as a content indicator

In the study of output screening, it was noted that titles and tracings are frequently inadequate in indicating the content of articles in the MEDLARS data base. The implication is that, although 58% of all the articles retrieved by MEDLARS are judged "of value" by requesters, by no means all of these articles are recognized as being potentially valuable when they appear as citations in demand search bibliographies. In the light of this, the requirement for including abstracts in the next-generation MEDLARS (as recognized in the Functional System Specifications for the National Library of Medicine, July 1, 1967) appears well-justified. In connection with this, we estimate that about 90% of input articles contain a usable content indicator in the form of abstract, summary or conclusions, although not all of these are in English.

Continuous quality control of the MEDLARS operation

A large-scale evaluation, of the type that has been undertaken, is useful in exposing the general weaknesses of the system. Such a study will also bring to light specific indexing failures, specific searching failures, and specific inadequacies of the index language. However, these specific failures must be regarded merely as symptomatic of kinds of failures that occur. A single evaluation study, however comprehensive, cannot be expected to discover more than a very small fraction of the specific inadequacies of the system. For example, we know that it is very difficult, if not impossible, to conduct a successful search on premature rupture of the fetal membranes, or one on gallbladder perforation. However, there are undoubtedly many other legitimate topics upon which MEDLARS cannot conduct a successful search, even though relevant literature exists in the system. Such specific inadequacies can only be discovered through continous monitoring of the MEDLARS operation.

We recommend that the Library, having concluded a large-scale study of the MEDLARS performance, should now investigate the feasibility of implementing procedures for the "continuous quality control" of MEDLARS operations. We recognize that continuous quality control is likely to be much more difficult to implement than a one-time evaluation. Nevertheless we feel that continuous system monitoring is ultimately essential to the success of any large retrieval system.

We visualize that "continuous quality control" would embrace at least the following functions:

1. Recognizing a request, within the scope of the system, that cannot adequately be conducted because of present indexing policies or vocabulary inadequacies. Any such requirements that are legitimate, and likely to be recurrent, indicate the need for changes in vocabulary or indexing policy.

2. Recognizing searches that have failed through defective interaction with the requester, poor searching strategies, vocabulary inadequacies, or indexing policies. Recall failures must be recognized by members of the MEDLARS staff, using similar methods to those employed in the present investigation (a heavy reliance would probably be put on the requester's own "known relevant" articles for this purpose). Precision failures must be identified primarily on the basis of feedback from the requester himself, and the present MEDLARS search appraisal form should be re-designed for this purpose.

Searches known to have performed badly, either in recall or precision, will require analysis to determine cause of failure. Such search analyses will be essential inputs to vocabulary control procedures, to decisions relating to indexing policy, and to search training functions.

3. Recognizing, in the indexing operation, items of subject matter that cannot be specifically expressed by present MeSH terms, and for which no terms exist in the entry vocabulary. The articles thus affected will require "flagging" by the indexer concerned, and subsequent action by the MeSH group. This action will involve the creation of a new MeSH term, a new provisional heading, or a new reference in the entry vocabulary.

Future use of the MEDLARS test corpus

During the conduct of this evaluation we have accumulated a corpus (of articles, indexing records, requests, searching strategies, and relevance assessments) that can be used for further analysis and experimentation. This corpus is already being drawn upon for a number of purposes, including the conduct of "search workshops" and the comparison of searching strategies prepared by various MEDLARS centers.

We recommend that this corpus should be the basis of further experimentation within MEDLARS. It would, for example, be a most valuable corpus upon which to conduct experiments on automatic indexing. In fact, a small part of it (18 searches and 276 documents) is already being used by Salton, at Cornell University, in the further testing of the SMART system. Natural language, free-text searching of abstracts would be another area of study, well worth investigating, for which the test corpus would be admirably suited (we have real requests and real relevance assessments). Finally, we recommend that the corpus be used for further studies on possible alternative modes of searching the MEDLARS data base. In particular, because many requesters can cite relevant articles at the

time they request a MEDLARS search, we suggest that NLM investigate the feasibility of deriving searching strategies automatically, by computer, on the basis of index terms assigned to articles known to be relevant to MEDLARS requests.

REFERENCES

1. National Library of Medicine. Description and History of MEDLARS. Bethesda, Maryland, National Library of Medicine, 1968. (In press)

2. Cleverdon, C. W. Evaluation of operational information retrieval systems. Part I. Identification of criteria. Cranfield, England, The College of Aeronautics, 1964.

3. Rodgers, D. J. A study of inter-indexer consistency. Washington, D. C., General Electric Company, 1961.

4. Hooper, R. S. Indexer consistency tests - origin, measurements, results and utilization. Bethesda, Maryland, IBM Corporation, 1965.

5. Lancaster, F. W. "Some observations on the performance of EJC role indicators in a mechanized retrieval system", Special Libraries, vol. 55, no. 10, 1964, pp. 696-701.

PART 4

APPENDICES

Appendix 1

Specimen page from Medical Subject Headings.

Specimen page from tree structure.

List of subheadings in use in 1967.

Specimen page from a "demand search bibliography".

Specimen search formulation.

Specimen "demand search request form".

SULFUR ISOTOPES (D1)

SULFURIC ACID (D1)

SUNBURN (C14)

SUNLIGHT (G3)
XU HELIOTHERAPY (G3)
XR LIGHT (H)

SUNSTROKE (C14)

SUPERFETATION (G1)

SUPERSTITIONS (I)

SUPPOSITORIES (D13)

SUPPURATION (C1, C16)
X PUS (C1, C16)

SUPRASELLAR CYST see under CRANIOPHARYNGIOMA (C2)

SURAMIN (D3)
X NAPHURIDE (D3)

SURFACE-ACTIVE AGENTS (D13)
XU TRITONS (D13)
XU TWEENS (D13)
XU WETTING AGENTS (D13)
XR ANTI-INFECTIVE AGENTS, LOCAL (D3)

SURFACE TENSION (H)

SURGERY (E4, G2)

SURGERY, MILITARY see under MILITARY MEDICINE (G2)

SURGERY, MILITARY see under WOUNDS AND INJURIES (C14)

SURGERY, MINOR (E4)

SURGERY, OPERATIVE (E4)
see also related
OPERATING ROOMS (G3)
POSTOPERATIVE COMPLICATIONS (C16)

SURGERY, ORAL (E4, E6, G2)
XU TUBERPLASTY (E4, E6, G2)
XU VESTIBULOPLASTY (E4, E6, G2)

SURGERY, PLASTIC (E4)

SURGERY, VETERINARY see under VETERINARY MEDICINE (G2)

SURGICAL DIATHERMY see under ELECTROCOAGULATION (E2)

SURGICAL EQUIPMENT (E4, E5)

SURGICAL INSTRUMENTS (E4)

SURGICAL INSURANCE see under INSURANCE, HEALTH (I)

SURGICAL WOUND DEHISCENCE (C16)

SURGICAL WOUND INFECTION (C1, C14, C16)

SURITAL see THIAMYLAL (D6)

SURVIVAL (I)
XR MILITARY MEDICINE (G2)

SUSPENSIONS (D13)

SUSTAINED-RELEASE PREPARATIONS see DELAYED-ACTION PREPARATIONS (D13)

SUTURE TECHNICS (E4)

SUTURES (E4)

SV40 VIRUS (B4)
X SIMIAN VIRUS 40 (B4)
X VACUOLATING AGENT (B4)

SWALLOWING see DEGLUTITION (G1)

SWEAT (A12)
XU PERSPIRATION (A12)

SWEAT GLANDS (A1)

SWEATING (G1)
XU DYSHIDROSIS (G1)
XU HYPERHIDROSIS (G1)
XU PERSPIRATION (G1)
XR CYSTIC FIBROSIS (C4, C5, C16)

SWEETENING AGENTS (D13)
XU CYCLAMATE (D13)
XU SACCHARIN (D13)

SWIFT'S DISEASE see ACRODYNIA (C10)

SWIMMING (E2)

SWIMMING POOLS (G3)

SWINE (B2)
XU HOGS (B2)

SWINE DISEASES (C15)

SWINE ERYSIPELAS see under ERYSIPELOID (C1, C15)

SYCOSIS (C1, C12)

SYMBIOSIS (G1)

SYMBIOTES (B3)

SYMBOLISM (F)

SYMPATHECTOMY (E4)

SYMPATHETIC GANGLIA see under GANGLIA, AUTONOMIC (A8)

A9 - SENSE ORGANS

SENSE ORGANS (NON MESH)	A9.			
EAR	A9.22.	A1.72.26.		
EAR, EXTERNAL	A9.22.16.	A1.72.39.		
EAR CANAL	A9.22.16.1			
EAR, MIDDLE	A9.22.32.			
EAR OSSICLES	A9.22.32.1			
EUSTACHIAN TUBE	A9.22.32.1			
TENSOR TYMPANI	A9.22.32.1	A2.72.62.		
TYMPANIC MEMBRANE	A9.22.32.1			
LABYRINTH	A9.22.48.			
COCHLEA	A9.22.48.1			
ORGAN OF CORTI	A9.22.48.1	A8.75.32.1		
SEMICIRCULAR CANALS	A9.22.48.1			
VESTIBULAR APPARATUS	A9.22.48.1			
EYE	A9.44.	A1.72.52.1		
ANTERIOR CHAMBER	A9.44.4.			
AQUEOUS HUMOR	A9.44.4.1	A12.13.12.		
CONJUNCTIVA	A9.44.9.			
CORNEA	A9.44.14.			
DESCEMET'S MEMBRANE	A9.44.14.1			
EYEBROWS	A9.44.19.	A1.72.52.1		
EYELASHES	A9.44.24.	A1.72.52.1		
EYELIDS	A9.44.29.	A1.72.52.1		
LACRIMAL APPARATUS	A9.44.34.			
LENS, CRYSTALLINE	A9.44.39.			
RETINA	A9.44.44.			
FUNDUS OCULI	A9.44.44.1			
MACULA LUTEA	A9.44.44.1			
RODS AND CONES	A9.44.44.1	A8.75.32.1		
SCLERA	A9.44.49.			
UVEA	A9.44.54.			
CHOROID	A9.44.54.1			
CILIARY BODY	A9.44.54.1			
IRIS	A9.44.54.1			
PUPIL	A9.44.54.1			
TRABECULAR MESHWORK (EYE)	A9.44.54.1			
VITREOUS BODY	A9.44.59.			
NASAL MUCOSA	A9.66.	A4.60.13.	A4.60.39.1	A10.88.48.
TASTE BUDS	A9.88.	A3.54.42.1		

Specimen page from tree structure

INDEX MEDICUS SUBHEADINGS - 1967

Alphabetical Listing

*abnormalities (A)
*administration & dosage (D)
*adverse effects (D,E,H)
*analysis (A,B,D)
*anatomy & histology (A,B)
*biosynthesis (D)
*blood (C,D,F)
*blood supply (A)
*cerebrospinal fluid (C,D,F)
*chemically induced (C,F)
*classification (B,C,D,E,F,G,H,I)
*complications (C,F)
*congenital (C)
*cytology (A,B)
*diagnosis (C,F)
*diagnostic use (D)
*drug effects (A,B,F,G)
*drug therapy (C,F)
*education (F,G,H,I)
*embryology (A,B)
*enzymology (A,B)
*etiology (C,F)
*familial & genetic (C,F)
*growth & development (A,B)
*history (C,D,E,F,G,H,I,J,K,L,M)
*immunology (B,C)
*injuries (A)
*innervation (A)
*instrumentation (E,F,G,H)
*isolation & purification (B)
*metabolism (A,B,C,D,F)
*microbiology (A,C)
*mortality (C,E,F)
*nursing (C,F)
*occurrence (C,F)
*pathogenicity (B)
*pathology (A,C,F)
*pharmacodynamics (D)
*physiology (A,B,D,G)
*physiopathology (A,C,F)
*poisoning (D)
*prevention & control (C,F)
*radiation effects (A,B,G)
*radiography (A,C)
*radiotherapy (C,F)
*rehabilitation (C,F)
*surgery (A,C,F)
*therapeutic use (D,H)
*therapy (C,F)
*toxicity (D)
*transplantation (A)
*urine (C,D,F)
*veterinary (C,E)

B. SUBTILIS AND BACTERIOPHAGE, TRANSDUCTION, GENETICS. 6=BACTERIOPHAGE. 5=BIOCHEMICAL GENETICS, GENES, CHROMOSOMES DNA, OR RNA. 4=GENETICS OR RADIATION GENETICS.

ABEL P, TRAUTNER TA
FORMATION OF AN ANIMAL VIRUS WITHIN A BACTERIUM.
Z VERERBUNGSL 95:66-72, 10 APR 64
*BACILLUS SUBTILIS, *BACTERIOPHAGE, *DNA, BACTERIAL, *DNA, VIRAL, EXPERIMENTAL LAB STUDY (4), GENETIC CODE (3), *VACCINIA VIRUS

APOSHIAN HV
A DTMPASE FOUND AFTER INFECTION OF BACILLUS SUBTILIS WITH PHAGE SP5C.
BIOCHEM BIOPHYS RES COMMUN 18:230-5, 18 JAN 65
*BACILLUS SUBTILIS, *BACTERIOPHAGE, EXPERIMENTAL LAB STUDY (4), METABOLISM, *NUCLEOTIDASES, NUCLEOTIDES, PHOSPHORUS ISOTOPES, RADIOMETRY

BARAT M, ANAGNOSTOPOULOS C, SCHNEIDER AM
LINKAGE RELATIONSHIPS OF GENES CONTROLLING ISOLEUCINE, VALINE, AND LEUCINE BIOSYNTHESIS IN BACILLUS SUBTILIS.
J BACT 90:357-69, AUG 65
AMINO ACID METABOLISM, AMINOHYDROLASES, ARGININE, *BACILLUS SUBTILIS, BACTERIOPHAGE, CHROMOSOME MAPPING (3), DNA, BACTERIAL, EXPERIMENTAL LAB STUDY (4), *GENES, HYDRO-LYASES, *ISOLEUCINE, *LEUCINE, LYSINE, METHIONINE, MUTATION, OXIDOREDUCTASES, PHENYLALANINE, THREONINE, ULTRAVIOLET RAYS, *VALINE

BAYREUTHER KE, ROMIG WR
POLYOMA VIRUS: PRODUCTION IN BACILLUS SUBTILIS.
SCIENCE 146:778-9, 6 NOV 64
*BACILLUS SUBTILIS, *BACTERIOPHAGE, *DNA, VIRAL, EXPERIMENTAL LAB STUDY (4), GENETIC CODE (3), *POLYOMA VIRUS

BOTT K, STRAUSS B
THE CARRIER STATE OF BACILLUS SUBTILIS INFECTED WITH THE TRANSDUCING BACTERIOPHAGE SP10.
VIROLOGY 25:212-25, FEB 65
*BACILLUS SUBTILIS, BACTERIOLYSIS, *BACTERIOPHAGE, CULTURE MEDIA, *DNA, BACTERIAL, *DNA, VIRAL, EXPERIMENTAL LAB STUDY (4), IMMUNE SERUMS, NUCLEOTIDYL TRANSFERASES, RNA, BACTERIAL, SPORES (3)

Specimen page from a
"demand search bibliography"

REQUEST NO. 250080

DEMAND SEARCH FORMULATION RECORD

DATE 1/16/67

TITLE

B. subtilis and bacteriophage or transduction.

11-14 ELEM. SYMBOL	17 EXPL.	18 LEVEL	19-31 CATEGORY NUMBER	ELEMENTS	TALLY
M1				Bacillus subtilis	
M5				Bacteriophage	
M6				Bacteriophage typing	
M10				Genetics	
M11				Radiation genetics	
M12				Bacteriolysis	
M13				Biochemical genetics	
M14				Genes	
M15				Chromosome mapping	
M16				Chromosomes	
M17				Mutation	
M18				Mutagens	
M19				DNA, bacterial	
M20				RNA, bacterial	
M21				DNA, viral	
M22				RNA, viral	
M23				Genetic code	

11-14 ELEM. SYMBOL	17 EXPL.	18 LEVEL	19-31 CATEGORY NUMBER	ELEMENTS	TALLY

ELEMENTS A, J, I, N, Y, X, AND SUMMATIONS

SECT. 7 3-	ELEM. SYMBOL 11-14	33-36	37-40	41-44	45-48	49-56
	M30	SUM	M5	M23		
	M31	SUM	M5	M6		
	M32	SUM	M12	M23		

SECT. 7 3-	ELEM. SYMBOL 11-14	33-36	37-40	41-44	45-48	49-56

REQUEST STATEMENTS

SECT. 7	ELEM. SYMBOL 9-10	11-80 COLUMNS	FOUND
4	1	M1 * M30.	
5	1	M31 + M32.	
6	1	M31	

BATCH NO.	DS MODULE	COMMENTS	RG MODULE	COMMENTS

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
NATIONAL LIBRARY OF MEDICINE

MEDLARS SEARCH REQUEST

DATE: 26 Sept 66

1. INDIVIDUAL WHO WILL ACTUALLY USE THE BIBLIOGRAPHY

Merlin G. Otteman MD USNR MC

TITLE

LCDR

ORGANIZATION

U. S. Naval Hospital Bethesda, Md NNMC

ADDRESS

U. S. Naval Hospital, Bethesda, Md. Dept. of Surgery

2. REQUEST SUBMITTED BY *(If different from above):*

Mary A. Dixon, Librarian, Edward Rhodes Stitt Library

3. DETAILED STATEMENT OF REQUIREMENTS *(Please be as specific as possible as to purpose, scope, definitions, limitations, etc.)*

A particularly long period of intestinal atony or ileus has been noted in patients following vagotomy and pyloroplasty in association with liver disease, chronic namely cirrhosis.

Am interested if there are any references noting this association.

4. TITLE OF PROJECT FOR WHICH SEARCH IS REQUESTED *(Omit if not applicable):*

5. MEDICAL TERMS PERTINENT TO REQUEST *(Optional).* DESCRIPTIONS CURRENTLY USED IN MEDLARS ARE PUBLISHED IN Medical Subject Headings, Part 2 of the JANUARY Issue of INDEX MEDICUS.

Vagotomy
Intestinal Ileus, Postoperative
Liver cirrhosis
Liver Diseases

6. LIMIT LANGUAGES TO
- [X] ACCEPT ALL
- [ ] ENGLISH
- [ ] FOREIGN *(Specify):*

7. PRINT SPECIFICATIONS:
- [ ] 3"X 5" CARDS
- [X] PAPER

PHS - 4667-1 (REV. 5-66)

Appendix 2

Sample of covering letter used to transmit random sample of articles.

Copy of Notes on Form for Document Evaluation.

Form to evaluate timeliness of the MEDLARS service.

Sample of completed Revised Statement of MEDLARS Request.

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
8600 ROCKVILLE PIKE
BETHESDA, MD., 20014

NATIONAL LIBRARY OF MEDICINE

REFER TO: NLM-ISD

October 11, 1966

Richard G. Buckles
Guest Scientist
Naval Medical Research Institute
National Naval Medical Center
Bethesda, Md. 20014

Dear Mr. Buckles

Enclosed you will find a set of journal articles of possible relevance to your recent request relating to: decompression sickness.

This set is composed largely of articles retrieved in the MEDLARS search for your request. However, certain articles found by supplementary searches may also have been included. Please carefully review the articles and complete for each the attached Form for Document Evaluation. You may, of course, keep the articles themselves for your own file. Please read the brief Notes on Form for Document Evaluation before completing these forms.

When you have reviewed all of the articles delivered to you, would you please give your evaluation of the timeliness of the MEDLARS service, and, if possible, a revised statement of your request, returning these along with your completed evaluation forms.

We are indeed most grateful for your assistance in the conduct of this important study, which has as its object the improvement of our literature searching service to the biomedical community.

Sincerely,

F. Wilfrid Lancaster
Information Systems Evaluator
Information Systems Division
National Library of Medicine

Enclosures

Notes on Form for Document Evaluation

In Item 1 on the Form, we are interested in discovering whether or not you were aware of the existence of the article before obtaining the MEDLARS search results. If you were, how did you become aware of its existence? It may be, for example, that the paper was sent to you by the author himself, that you regularly scan the journal in which it was published, that it was drawn to your attention by a colleague, or that it was found through a prior search of the literature. We would like to know what particular channel brought this document to your attention.

Item 2 asks you to grade each document, in relation to the information need that prompted your request to MEDLARS, on the scale "of major value", "of minor value", "of no value", and to indicate reasons for making these evaluations. If you were able to name a number of relevant articles prior to the MEDLARS search, please make your present assessments in line with your earlier gradings (see attached Record of Known Relevant Documents). A "major value" article in this sample should be roughly equivalent in value to a "major value" article named in advance of search. Likewise for "minor value" documents.

It is particularly important that we obtain precise reasons for your judging an article "of no value". If you discover, for example, that the article does not relate to the subject of your information need, please indicate in what way the subject matter differs from that desired. If the mode of treatment differs from that needed (e.g. clinical rather than experimental or vice versa) please so indicate. In other cases you may doubt the value or validity of the work described. Whatever your reason for determining that an article is "of no value", please record this under 2(c).

2(d) should be checked in those cases in which you are unable to make a value judgment on the article, in relation to your information need, because of lack of familiarity with the language in which it is written. In these cases, we are interested in knowing whether or not you intend to take any steps (such as consulting an English abstract or having a complete or partial translation made) to learn the contents of the article. If you do not intend to go further with this article, we should like to know why. Is it that you doubt the value of work carried out in that particular field within the country concerned? Or do you feel that you have sufficient material on your topic from other sources to make a translation of this article unnecessary? You will probably want to review the entire set of citations retrieved before answering this question.

NATIONAL LIBRARY OF MEDICINE
Bethesda, Maryland

MEDLARS EVALUATION PROJECT

Request No:_____

Timeliness of MEDLARS Service

Did delay in MEDLARS response to your request significantly reduce the value of the search?

Yes ☐ In what way? (Please explain)

No ☐

REVISED STATEMENT OF MEDLARS REQUEST

You have now had the opportunity to examine a sample of articles retrieved in response to your recent MEDLARS request. It would be very helpful in our evaluation if, having seen this sample, you would now re-phrase more precisely the wording of your original request:

Too many clinical papers as a result of the original request. Possibly I should have added the words 'experimental lab studies'. Or a number of more defined searches might have been better, such as

1. Lens development and biosynthesis of protein, RNA and DNA
2. Experimental lab studies on eye ~~growth~~ embryology, growth and development.

If you do not feel that the original wording needs changing, please check this box: ☐

Thank you for your continued cooperation.

F. Wilfrid Lancaster
Information Systems Evaluator
Information Systems Division
National Library of Medicine

3 Lens capsule and basement membrane (embryology of)

Appendix 3

Sample of a completed "analysis worksheet" as prepared during the analysis of Search #539. Similar worksheets were prepared during the analysis of each of the 302 searches.

SEARCH # 539

RECALL

	MAJOR	MINOR	TOTAL	MAJOR RECALL	MINOR RECALL	TOTAL RECALL
1. KNOWN BY REQUESTER		11	11		$\frac{8}{11} = 72.7\%$	$\frac{8}{11} = 72.7\%$
2. KNOWN BY REQUESTER NOT IM (INCLUDED IN 1.)		3	3		$\frac{2}{3} = 66.7\%$	$\frac{2}{3} = 66.7\%$
3. FOUND BY NLM, JUDGED RELEVANT						
4. 2 and 3 COMBINED			AS 2.			
5. GRAND TOTAL		11	11		$\frac{8}{11} = 72.7\%$	$\frac{8}{11} = 72.7\%$

Recall for Section 5:

$\frac{8}{11} = 72.7\%$

SEARCH # 539

PRECISION

TOTAL RETRIEVED	EXPECTED	SAMPLE	ASSESSED	RELEVANT	PRECISION RATIO	MAJOR ONLY
52 Section 5 : 43	51-200	19	18	11	$\frac{11}{18} = 61.1\%$ Section 5 $\frac{10}{15} = 66.7\%$	$\frac{0}{18} = 0$

	MAJOR	MINOR	NO VALUE	CANNOT ASSESS
Known Prior to MEDLARS Search		8		
Not Known Prior to MEDLARS Search		3	7	1

REQUEST Auditory monitoring of speech using the technique of delayed auditory feedback.

SEARCH FORMULATION

[ACOUSTICS
AUDIOMETRY
AUDITORY PERCEPTION
HEARING
HEARING TESTS
NOISE
PHONETICS
SOUND
SPEECH
SPEECH DISORDERS
STUTTERING
TAPE RECORDING
VOICE
VOICE TRAINING
SPEECH THERAPY
HEARING DISORDERS]

AND [CHRONAXY
CYBERNETICS
FEEDBACK
[MODELS
NEURAL
ANALYZERS] AND [TIME
REACTION
TIME]]

TIME FACTORS

SEARCH REQUESTED: 3/22

FORMULATED 4/5

DS MODULE: 4/7

REFORMULATED:

DS MODULE:

RG MODULE: 4/7

DELIVERED: 4/10

VALUE: Not affected

No evidence of reformulation

REQUEST CHARACTERISTICS

ORGANIZATION

ACADEMIC

SUBJECT

TECHNICS

INTERACTION

LOCAL – POSITIVE

NLM 12

FOREIGN LANGUAGE MATERIAL

LANGUAGE RESTRICTION: None

TOTAL SAMPLE OF 19:

13 ENGLISH
6 FOREIGN

LANGUAGE	NUMBER OF ITEMS	ASSESSED	RELEVANT	ACTION ON UNASSESSED ITEMS
GER	5	5	2	None: unlikely to be of value
CZ	1	0		

FOREIGN ASSESSED: $\frac{5}{6}$

PRECISION RATIO: $\frac{2}{5}$

PREVIOUS AWARENESS

PREVIOUSLY AWARE OF 8 OF THE 19 ITEMS

BECAME AWARE BECAUSE:

Found in Index Medicus : 8

	RETRIEVED AND OF VALUE	KNOWN PREVIOUSLY	NOVELTY RATIO
MAJOR			
MINOR	11	8	3/11
TOTAL	11	8	3/11

INDEXING EXHAUSTIVITY

(recognizing IM terms only)

RECALL ⟶

PRECISION ⟶

EFFECT OF REVISION OF SEARCH FORMULATION

RECALL ⟶

PRECISION ⟶

EFFECT OF ~~Ra-Rb-Rc~~ 4-5-6 AS RANKING DEVICE

Complete Search	Total Retrieved	Recall Ratio	Precision Ratio
Section 5 ~~Rb~~ only	43	$\frac{8}{11} = 72.7\%$	$\frac{11}{18} = 61.1\%$
~~Rc only~~		$\frac{8}{11} = 72.7\%$	$\frac{10}{15} = 66.7\%$
~~Rb and Rc~~ Combined			

COVERAGE

Of the 15 articles, within the time span of MEDLARS, named by the requester, 11 were shown to be in the data base.

Missing items:

Psychonomic Science (3 articles):
journal not indexed

J Speech Hear Res — 1966
issue — not indexed into
base as of April 1967

SEARCH NO. 539 REQUESTER: FARGO

NUMBER	JOURNAL	DATE	LANGUAGE	SET	KNOWN BEFORE	DECISION
539/1	Arch Ohr Nas Kehlkopfheilk	12/63	GER	4	N	NV
539/2	ditto	11/63	GER	4	N	NV
539/3	Cesk Otolaryng	4/64	CZ	4	N	U
539/4	Percept Motor Skills	10/65		5	N	MINOR
539/5	ditto	8/64		5	Y	MINOR
539/6	Mschr. Ohrenheilk	7/64	GER	5	N	NV
539/7	J Psycrol	1/65		5	Y	MINOR
539/8	J Laryng	7/64		5	N	NV
539/9	IEEE Trans Biomed Electronics	1-4/64		5	N	NV
539/10	H N O	7/64	GER	5	Y	MINOR
539/11	Folia Phoniat (Basel)	66		5	N	MINOR
539/12	Psycrol Rev	7/65		5	N	NV
539/13	Quart J Exp Psycrol	5/65		5	Y	MINOR
539/14	Amer J Ment Defic	3/65		5	N	NV
539/15	Acta Neur Sup (Fraha)	65		5	Y	MINOR
539/16	ditto	65		5	Y	MINOR
539/17	J Speech Hearing Dis	8/64		5	Y	MINOR
539/18	Progr Biocybern	64		4	N	MINOR
539/19	Z Psycrol	1/65	GER	5	Y	MINOR

*Indicates item of some interest in relation to some other need or project.

3

REASONS FOR NONRETRIEVAL OF WANTED ITEMS

(3 minor value)

CAUSE	NUMBER OF MISSED RETRIEVALS AFFECTED
Indexing	
Lack of specificity	1
Omitted important term	1
Computer processing	1

EXPLANATION

Article on effect of drugs on performance under delayed auditory feedback — deals with verbal, arithmetic and color tests — but indexed only under the general PSYCHOLOGICAL TESTS (non-depth journal: 4 terms assigned).

Article on "computer systems control of delayed auditory feedback": indexed under HEARING and SPEECH but not under FEEDBACK (term appears in title).

The third item matters the search formulation exactly — should have been retrieved on FEEDBACK and [SPEECH or HEARING]. Failure must be due to some aspect of machine processing

7

REASONS FOR RETRIEVAL OF UNWANTED ITEMS

CAUSE	NUMBER OF UNWANTED RETRIEVALS AFFECTED
Index language Lack of specificity	7

EXPLANATION

It is not possible to express precisely the notion of "delayed auditory feedback". FEEDBACK and HEARING retrieved fine articles on feedback mechanisms in the physiology of hearing (i.e., the "acoustic reflex feedback loop")

Moreover, the term FEEDBACK was not available in 1964. Presumably for this earlier material, the searcher used the term CYBERNETICS. CYBERNETICS and HEARING TESTS or HEARING retrieved two relevant articles: hearing in information theory; binaural hearing during noise.

Appendix 4

Complete set of recall and precision ratios for the 302 searches.

Appendix 4

COMPLETE SET OF RECALL AND PRECISION RATIOS FOR THE 302 SEARCHES

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
1	344	19/24= 79.2%	6/24= 25%	15/17= 88.2%	5/7= 71.4%	As Total		As Total	As Total
2	11	8/9= 88.9%	2/9= 22.2%	3/4= 75%	2/3= 66.7%	As Total		2/2= 100%	1/1= 100%
3	209	17/22= 77.3%	11/22= 50%	11/12= 91.7%	9/9= 100%	4/4= 100%	7/8= 87.5%	As Total	As Total
4	32	18/25= 72%	11/25= 44%	1/4= 25%	1/3= 33.3%	0/1= 0*	As Total*	As Total	As Total
5	22	17/19= 89.5%	12/19= 63.2%	2/8= 25%	2/7= 28.6%	As Total		As Total	As Total
6	30	9/29= 31%	5/29= 17.2%	5/7= 71.4%	5/6= 83.3%	4/4= 100%	2/4= 50%	As Total	As Total
7	42	5/22= 22.7%	2/22= 9.1%	5/7= 71.4%	3/4= 75%	As Total		As Total	As Total
8	300 (1235)**	1/18= 5.6%	0/18= 0	0/15= 0		As Total		As Total	
9	860	17/23= 73.9%	7/23= 30.4%	25/25= 100%	12/12= 100%	20/20= 100%	5/5= 100%	5/5= 100%	1/1= 100%
10	301	15/24= 62.5%	5/24= 20.8%	3/3= 100%	3/3= 100%	As Total		As Total	As Total

* The recall base component "Known by the requester" and the component found manually need not be mutually exclusive (i.e., they may overlap). In search #4, for example, the only item known by the requester was also among the four found independently by manual search.

** Where the "machine search" result (i.e., actual number of citations matching the search formulation) is different from the number of citations actually submitted to the requester (because of post-editing of printout or the decision to arbitrarily set a limit on the number of citations to be printed), the former figure is presented in parentheses in this column.

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
11	1167 (1888)	0/26= 0	0/26= 0	1/2= 50%	1/2= 50%	1/1= 100%	0/1= 0	As Total	As Total
13	70	14/21= 66.7%	6/21= 28.6%	7/7= 100%	6/6= 100%	2/2= 100%	5/5= 100%	As Total	As Total
14	17	6/12= 50%	3/12= 25%	1/3= 33.3%	1/3= 33.3%	1/2= 50%	1/2= 50%	1/2= 50%	1/2= 50%
16	198	17/25= 68%	5/25= 20%	11/16= 68.7%	11/16= 68.7%	As Total		As Total	As Total
17	216	12/29= 41.4%	7/29= 24.1%	0/3= 0		As Total		As Total	
18	43	7/11= 63.6%	2/11= 18.2%	0/4= 0	0/4= 0	As Total	0/1= 0	As Total	As Total
19	40	5/28= 17.9%	3/28= 10.7%	2/4= 50%		None Supplied	As Total	As Total	
20	357	8/21= 38.1%	3/21= 14.3%	8/14= 57.1%	6/11= 54.5%	As Total		5/10= 50%	4/8= 50%
22	500 (775)	4/22= 18.2%	2/22= 9.1%	4/6= 66.7%	4/5= 80%	2/4= 50%	2/2= 100%	3/5= 60%	3/4= 75%
24	585	4/17= 23.5%	0/17= 0	3/5= 60%	1/2= 50%	1/1= 100%	2/4= 50%	As Total	As Total
25	61	7/22= 31.8%	4/22= 18.2%	3/4= 75%	3/3= 100%	0/1= 0	3/3= 100%	As Total	As Total
27	62	17/25= 68%	10/25= 40%	4/14= 28.6%	4/9= 44.4%	4/6= 66.7%	0/8= 0	1/9= 11.1%	1/6= 16.7%
29	21	16/19= 84.2%	9/19= 47.4%	5/7= 71.4%	3/5= 60%	2/3= 66.7%	3/4= 75%	4/5= 80%	2/3= 66.7%
30	57	4/17= 23.5%	1/17= 5.9%	0/1= 0	0/1= 0	None Known in Base	As Total	As Total	As Total
32	36	21/29= 72.4%	8/29= 27.6%	2/6= 33.3%	2/5= 40%	0/2= 0	2/4= 50%	2/5= 40%	2/4= 50%
33	37	20/24= 83.3%	16/24= 66.7%	3/6= 50%	3/3= 100%	2/4= 50%	2/3= 66.7%	2/3= 66.7%	2/2= 100%
34	285	13/27= 48.1%	1/27= 3.7%	0/8= 0	0/7= 0	0/3= 0	0/5= 0	0/5= 0	0/4= 0
36	53	13/18= 72.2%	6/18= 33.3%	6/7= 85.7%	6/6= 100%	As Total		As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
38	492	13/20= 65%	7/20= 35%	2/4= 50%	2/2= 100%	None Supplied	As Total	As Total	As Total
39	42	15/27= 55.6%	3/27= 11.1%	2/4= 50%	1/1= 100%	1/1= 100%	1/3= 33.3%	As Total	As Total
40	22	5/17= 29.4%	2/17= 11.8%	1/6= 16.7%	1/6= 16.7%	0/3= 0	1/4= 25%	1/5= 20%	1/5= 20%
42	92	15/25= 60%	5/25= 20%	6/7= 85.7%	4/5= 80%	1/1= 100%	5/6= 83.3%	As Total	As Total
43	173	14/21= 66.7%	8/21= 38.1%	2/4= 50%	2/4= 50%	2/3= 66.7%	0/1= 0	As Total	As Total
44	6	4/4= 100%	0/4= 0	0/9= 0	0/4= 0	0/7= 0	0/3= 0	0/3= 0	0/1= 0
45	533	3/17= 17.6%	1/17= 5.9%	5/6= 83.3%	5/6= 83.3%	4/4= 100%	1/2= 50%	As Total	As Total
46	339 (690)	2/10= 20%	0/10= 0	0/2= 0	0/2= 0	0/1= 0	0/1= 0	As Total	As Total
47	79	4/29= 13.8%	4/29= 13.8%	4/6= 66.7%	4/6= 66.7%	1/1= 100%	3/5= 60%	As Total	As Total
48	139	17/24= 70.8%	3/24= 12.5%	5/13= 38.5%	2/5= 40%	0/6= 0	5/7= 71.4%	As Total	As Total
49	2	0/2= 0	0/2= 0	0/1= 0		0/1= 0		As Total	
51	25	7/13= 53.8%	3/13= 23.1%	0/2= 0		None Known in Base	As Total	As Total	
52	5	3/5= 60%	1/5= 20%	2/2= 100%	1/1= 100%	1/1= 100%	1/1= 100%	As Total	As Total
53	764	10/27= 37%	1/27= 3.7%	5/5= 100%		None Supplied	As Total	As Total	
55	235	14/25= 56%	3/25= 12%	4/4= 100%	1/1= 100%	1/1= 100%	3/3= 100%	As Total	As Total
56	0	0/0= 100%	0/0= 100%	0/0= 100%	0/0= 100%	None Known	None Found	As Total	As Total
57	45	17/26= 65.4%	6/26= 23.1%	4/7= 57.1%	4/7= 57.1%	2/3= 66.7%	4/6= 66.7%	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
59	256	16/22= 72.7%	11/22= 50%	2/5= 40%	1/2= 50%	None Known in Base	As Total	As Total	As Total
60	286	12/27= 44.4%	5/27= 18.5%	12/15= 80%	9/12= 75%	6/9= 66.7%	6/7= 85.7%	As Total	As Total
61	213	7/19= 36.8%	3/19= 15.8%	4/8= 50%	4/8= 50%	As Total		As Total	As Total
62	22	17/20= 85%	10/20= 50%	1/1= 100%	1/1= 100%	None Supplied	As Total	As Total	As Total
63	240	13/24= 54.2%	10/24= 41.7%	3/5= 60%	3/5= 60%	None Supplied	As Total	As Total	As Total
64	156	14/27= 51.9%	4/27= 14.8%	10/13= 76.9%	7/9= 77.8%	As Total		As Total	As Total
65	224	5/28= 17.9%	1/28= 3.6%	5/6= 83.3%	3/3= 100%	3/4= 75%	2/2= 100%	As Total	As Total
67	11	2/9= 22.2%	2/9= 22.2%	2/2= 100%	2/2= 100%	1/1= 100%	1/1= 100%	As Total	As Total
68	29	12/28= 42.9%	4/28= 14.3%	6/9= 66.7%	2/3= 66.7%	1/2= 50%	5/8= 62.5%	As Total	As Total
70	89	11/23= 47.8%	5/23= 21.7%	1/3= 33.3%	1/2= 50%	0/1= 0	1/2= 50%	1/2= 50%	1/1= 100%
71	283	4/8= 50%	3/8= 37.5%	12/14= 85.7%	10/12= 83.3%	5/6= 83.3%	7/8= 87.5%	As Total	As Total
72	264 (1398)	17/29= 58.6%	9/29= 31%	9/9= 100%	2/2= 100%	1/1= 100%	8/8= 100%	As Total	As Total
73	140	8/19= 42.1%	8/19= 42.1%	6/7= 85.7%	3/4= 75%	3/3=100%	3/4= 75%	As Total	As Total
76	104	8/23= 34.8%	0/23= 0	2/8= 25%	0/2= 0	As Total		2/6= 33.3%	0/1= 0
77	294	11/25= 44%	4/25= 16%	10/15= 66.7%	4/8= 50%	3/6= 50%	7/9= 77.8%	7/9= 77.8%	4/6= 66.7%
78	529	10/21= 47.6%	6/21= 28.6%	10/19= 52.6%	5/11= 45.5%	7/15= 46.7%	3/4= 75%	6/8= 75%	3/3= 100%
79	17	11/17= 64.7%	5/17= 29.4%	3/6= 50%	2/4= 50%	2/5= 40%	1/1= 100%	As Total	As Total
80	115	13/25= 52%	7/25= 28%	1/4= 25%	1/3= 33.3%	None Supplied	As Total	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
82	315	10/27= 37%	5/27= 18.5%	13/20= 65%	12/19= 63.2%	9/16= 56.2%	6/6= 100%	6/6= 100%	5/5= 100%
83	23	1/15= 6.7%	1/15= 6.7%	0/1= 0		None Supplied	As Total	As Total	
84	437	5/22= 22.7%	2/22= 9.1%	1/1= 100%	1/1= 100%	None Supplied	As Total	As Total	As Total
90	38	21/24= 87.5%	13/24= 54.2%	5/5= 100%	5/5= 100%	1/1= 100%	4/4= 100%	4/4= 100%	4/4= 100%
91	777	6/27= 22.2%	0/27= 0	2/4= 50%		None Known in Base	As Total	As Total	
92	500 (1147)	17/23= 73.9%	7/23= 30.4%	7/12= 58.3%	4/8= 50%	2/5= 40%	5/8= 62.5%	5/10= 50%	2/4= 50%
93	6	6/6= 100%	3/6= 50%	0/4= 0	0/1= 0	As Total		As Total	As Total
94	17	5/14= 35.7%	3/14= 21.4%	2/4= 50%	2/3= 66.7%	None Supplied	As Total	As Total	As Total
95	46	10/23= 43.5%	9/23= 39.1%	6/8= 75%	5/5= 100%	1/3= 33.3%	5/5= 100%	As Total	As Total
96	65	7/27= 25.9%	4/27= 14.8%	1/4= 25%	0/1= 0	1/1= 100%	0/3= 0	0/3= 0	0/1= 0
97	19	8/13= 61.5%	5/13= 38.5%	4/6= 66.7%	4/6= 66.7%	As Total		As Total	As Total
98	875	13/18= 72.2%	10/18= 55.6%	2/5= 40%	2/5= 40%	As Total		As Total	As Total
99	171	11/17= 64.7%	2/17= 11.8%	4/5= 80%	2/3= 66.7%	1/1= 100%	3/4= 75%	As Total	As Total
100	223	10/22= 45.5%	4/22= 18.2%	9/18= 50%	4/6= 66.7%	4/10= 40%	5/8= 62.5%	As Total	As Total
101	459	1/28= 3.6%	0/28= 0	3/11= 27.3%	1/4= 25%	As Total		As Total	As Total
102	99	2/25= 8%	1/25= 4%	6/11= 54.5%	2/6= 33.3%	2/6= 33.3%	4/5= 80%	As Total	As Total
103	87	6/23= 26.1%	1/23= 4.3%	2/3= 66.7%	2/2= 100%	1/1= 100%	2/3= 66.7%	As Total	As Total
104	16	2/15= 13.3%	1/15= 6.7%	1/1= 100%		None Known	As Total	As Total	

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
105	12	0/10= 0	0/10= 0	0/2= 0	0/2= 0	None Known	As Total	As Total	As Total
106	260	14/22= 63.6%	9/22= 40.9%	4/6= 66.7%	4/5= 80%	None Supplied	As Total	As Total	As Total
107	15	5/13= 38.5%	0/13= 0	1/1= 100%		None Supplied	As Total	As Total	
108	471	17/22= 77.3%	10/22= 45.5%	5/5= 100%	4/4= 100%	None Supplied	As Total	As Total	As Total
110	245	13/17= 76.5%	5/17= 29.4%	17/18= 94.4%	7/7= 100%	15/16= 93.7%	5/5= 100%	11/12= 91.7%	4/4= 100%
113	28	17/26= 65.4%	6/26= 23.1%	4/7= 57.1%	01= 0	3/5= 60%	2/3= 66.7%	2/3= 66.7%	
114	63	6/25= 24%	0/25= 0	1/2= 50%		None Known	As Total	As Total	
115	0	0/0= 0	0/0= 0	0/4= 0	0/2= 0	None Supplied	As Total	As Total	As Total
116	505	22/25= 88%	12/25= 48%	8/9= 88.9%	7/8= 87.5%	3/4= 75%	5/5= 100%	5/5= 100%	4/4= 100%
117	404	11/22= 50%	5/22= 22.7%	3/5= 60%	3/3= 100%	None Known in Base	As Total	As Total	As Total
118	500 (775)	5/26= 19.2%	4/26= 15.4%	2/4= 50%	2/4= 50%	As Total		As Total	As Total
120	340	16/23= 69.6%	1/23= 4.3%	2/3= 66.7%	2/2= 100%	None Supplied	As Total	As Total	As Total
121	313	11/29= 37.9%	3/29= 10.3%	19/27= 70.4%	16/20= 80%	14/22= 63.6%	5/5= 100%	As Total	As Total
122	70	22/22= 100%	11/22= 50%	6/6= 100%	5/5= 100%	1/1= 100%	6/6= 100%	As Total	As Total
123	33	12/23= 52.2%	4/23= 17.4%	2/9= 22.2%	0/4= 0	2/4= 50%	0/6= 0	1/8= 12.5%	0/4= 0
124	378	8/25= 32%	2/25= 8%	6/6= 100%	4/4= 100%	3/3= 100%	4/4= 100%	5/5= 100%	3/3= 100%
125	53	19/24= 79.2%	13/24= 54.2%	3/7= 42.9%	2/3= 66.7%	2/4= 50%	1/3= 33.3%	As Total	As Total
126	125	5/23= 21.7%	2/23= 8.7%	2/4= 50%	2/3= 66.7%	None Supplied	As Total	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
128	13	10/12= 83.3%	6/12= 50%	1/6= 16.7%	1/5= 20%	As Total		As Total	As Total
132	103	8/14= 57.1%	6/14= 42.9%	1/3= 33.3%	1/2= 50%	None Supplied	As Total	As Total	As Total
133	285	3/22= 13.6%	1/22= 4.5%	3/6= 50%	3/3= 100%	2/2= 100%	2/5= 40%	As Total	As Total
134	24	18/18= 100%	9/18= 50%	4/7= 57.1%	4/6= 66.7%	2/3= 66.7%	2/4= 50%	As Total	As Total
136	500 (705)	4/17= 23.5%	1/17= 5.9%	2/3= 66.7%	2/2= 100%	1/2= 50%	2/2= 100%	As Total	As Total
137	173	12/13= 92.3%	0/13= 0	4/5= 80%	1/1= 100%	None Supplied	As Total	As Total	As Total
147	19	8/16= 50%	2/16= 12.5%	1/4= 25%		None Known in Base	As Total	As Total	
148	500 (526)	4/27= 14.8%	0/27= 0	4/5= 80%	1/1= 100%	1/1= 100%	3/4= 75%	3/4= 75%	
149	79	8/25= 32%	4/25= 16%	9/9= 100%	8/8= 100%	6/6= 100%	8/8= 100%	As Total	As Total
151	17	5/11= 45.5%	1/11= 9.1%	No Recall Base		None Known	None Found		
152	103	12/17= 70.6%	7/17= 41.2%	3/11= 27.3%	3/11= 27.3%	1/2= 50%	2/9= 22.2%	2/9= 22.2%	2/9= 22.2%
153	539	11/23= 47.8%	0/23= 0	10/14= 71.4%	9/11= 81.8%	8/10= 80%	3/5= 60%	3/5= 60%	3/4= 75%
155	250	13/23= 56.5%	6/23= 26.1%	6/12= 50%	3/7= 42.9%	3/7= 42.9%	4/6= 66.7%	4/6= 66.7%	2/4= 50%
157	53	4/18= 22.2%	1/18= 5.6%	1/2= 50%	1/1= 100%	1/1= 100%	0/1= 0	0/1= 0	
158	9	9/9= 100%	7/9= 77.8%	1/2= 50%	1/2= 50%	None Supplied	As Total	As Total	As Total
159	31	22/29= 75.9%	9/29= 31%	2/3= 66.7%	1/1= 100%	None Supplied	As Total	As Total	As Total
160	6	5/6= 83.3%	4/6= 66.7%	1/9= 11.1%	1/6= 16.7%	None Supplied	As Total	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
162	14	7/12= 58.3%	5/12= 41.7%	2/10= 20%	1/5= 20%	2/5= 40%	0/5= 0	0/6= 0	0/4= 0
163	506	26/27= 96.3%	22/27= 81.5%	17/17= 100%	17/17= 100%	As Total		As Total	As Total
165	17	5/12= 41.7%	1/12= 8.3%	0/8= 0	0/2= 0	0/2= 0	0/6= 0	As Total	As Total
166	13	7/9= 77.8%	4/9= 44.4%	0/5= 0	0/5= 0	None Supplied	As Total	As Total	As Total
169	1	0/1= 0	0/1= 0	0/0= 100%	0/0= 100%	None Known in Base	None Found	As Total	As Total
171	33	33/33= 100%	29/33= 87.9%	8/14= 57.1%	8/9= 88.9%	0/5= 0	8/9= 88.9%	8/9= 88.9%	8/9= 88.9%
172	6	3/5= 60%	1/5= 20%	0/3= 0	0/2= 0	None Supplied	As Total	As Total	As Total
173	96	8/15= 53.3%	0/15= 0	2/3= 66.7%	1/1= 100%	None Supplied	As Total	As Total	As Total
174	103	10/20= 50%	2/20= 10%	3/11= 27.3%	1/1= 100%	2/7= 28.6%	1/4= 25%	1/4= 25%	1/1= 100%
175	17	9/12= 75%	6/12= 50%	2/2= 100%	2/2= 100%	None Known	As Total	As Total	As Total
176	123	2/12- 16.7%	0/12= 0	1/1= 100%	1/1= 100%	None Supplied	As Total	As Total	As Total
177	50	3/23= 13%	2/23= 8.7%	1/5= 20%	1/5= 20%	None Supplied	As Total	As Total	As Total
179	375	10/25= 40%	2/25= 8%	8/8= 100%	3/3= 100%	1/1= 100%	7/7= 100%	As Total	As Total
180	71	5/11= 45.5%	3/11= 27.3%	2/2= 100%	2/2= 100%	1/1= 100%	1/1= 100%	As Total	As Total
181	34	4/23= 17.4%	3/23= 13%	3/5= 60%	3/5= 60%	2/2= 100%	2/4= 50%	2/4= 50%	2/4= 50%
182	42	9/23= 39.1%	7/23= 30.4%	2/3= 66.7%	2/2= 100%	None Known in Base	As Total	As Total	As Total
183	27	16/24= 66.7%	9/24= 37.5%	3/6= 50%	1/2= 50%	2/4= 50%	2/3= 66.7%	2/3= 66.7%	1/2= 50%
185	250 (927)	2/16= 12.5%	0/16= 0	1/7= 14.3%	1/2= 50%	0/3= 0	1/4= 25%	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
187	157	13/17= 76.5%	9/17= 52.9%	3/3= 100%	3/3= 100%	1/1= 100%	2/2= 100%	As Total	As Total
188	9	9/9= 100%	9/9= 100%	3/8= 37.5%	3/7= 42.9%	None Known in Base	As Total	As Total	As Total
189	116	31/31= 100%	28/31= 90.3%	6/6= 100%	5/5= 100%	None Supplied	As Total	As Total	As Total
190	388 (628)	7/19= 36.8%	2/19= 10.5%	6/7= 85.7%	6/7= 85.7%	2/2= 100%	4/5= 80%	As Total	As Total
191	63	11/18= 61.1%	3/18= 16.7%	3/17= 17.6%	2/12= 16.7%	3/7= 42.9%	2/12= 16.7%	2/14= 14.3%	1/10= 10%
192	5	0/5= 0	0/5= 0	0/0= 100%	0/0= 100%	None Known	None Found	As Total	As Total
194	94	12/20= 60%	2/20= 10%	5/5= 100%	4/4= 100%	None Known in Base	As Total	As Total	As Total
195	112	14/18= 77.8%	4/18= 22.2%	3/6= 50%	3/6= 50%	1/1= 100%	2/5= 40%	As Total	As Total
197	87	8/22= 36.4%	2/22= 9.1%	1/1= 100%	1/1= 100%	None Supplied	As Total	As Total	As Total
198	142	15/17= 88.2%	7/17= 41.2%	7/8= 87.5%	5/6= 83.3%	None Known in Base	As Total	As Total	As Total
199	822	21/21= 100%	10/21= 47.6%	5/9= 55.6%	4/8= 50%	0/2= 0	5/7= 71.4%	As Total	As Total
200	96	10/21= 47.6%	8/21= 38.1%	7/8= 87.5%	6/6= 100%	3/3= 100%	6/7= 85.7%	As Total	As Total
201	500 (699)	3/27= 11.1%	0/27= 0	1/6= 16.7%	1/5= 20%	0/1= 0	1/5= 20%	As Total	As Total
202	33	7/21= 33.3%	2/21= 9.5%	1/4= 25%	0/1= 0	0/1= 100%	1/3= 33.3%	As Total	As Total
203	38	1/11= 9.1%	0/11= 0	0/3= 0	0/3= 0	0/1= 0	0/2= 0	As Total	As Total
206	259	20/22= 90.9%	15/22= 68.2%	5/8= 62.5%	4/7= 57.1%	0/3= 0	5/5= 100%	As Total	As Total
207	200	13/29= 44.8%	3/29= 10.3%	8/8= 100%	8/8= 100%	2/2= 100%	6/6= 100%	As Total	As Total
208	39	27/32= 84.4%	9/32= 28.1%	7/7= 100%	3/3= 100%	4/4= 100%	6/6= 100%	6/6= 100%	3/3= 100%

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
209	88	21/24= 87.5%	12/24= 50%	7/7= 100%	4/4= 100%	1/1= 100%	7/7= 100%	As Total	As Total
210	8 (307)	3/7= 42.9%	2/7= 28.6%	1/2= 50%	1/2= 50%	None Supplied	As Total	As Total	As Total
211	125	1/27= 3.7%	0/27= 0	1/1= 100%	1/1= 100%	As Total	None Found	As Total	As Total
212	278	4/12= 33.3%	2/12= 16.7%	11/11= 100%	7/7= 100%	9/9= 100%	4/4= 100%	5/5= 100%	2/2= 100%
213	1	0/1= 0	0/1= 0	0/5= 0	0/5= 0	As Total		As Total	As Total
214	273	5/26= 19.2%	1/26= 3.8%	10/11= 90.9%	7/7= 100%	5/5= 100%	5/6= 83.3%	9/10= 90%	6/6= 100%
215	174	14/24= 58.3%	5/24= 20.8%	5/6= 83.3%	5/6= 83.3%	As Total		As Total	As Total
216	44	6/28= 21.4%	1/28= 3.6%	1/1= 100%		None Known	As Total	As Total	
217	20	2/16= 12.5%	2/16= 12.5%	1/2= 50%	1/2= 50%	None Known	As Total	As Total	As Total
220	104	4/7= 57.1%	2/7= 28.6%	1/2= 50%	1/1= 100%	None Known in Base	As Total	As Total	As Total
221	191	3/27= 11.1%	2/27= 7.4%	1/2= 50%	1/2= 50%	None Known	As Total	As Total	As Total
224	500 (803)	14/22= 63.6%	9/22= 40.9%	6/8= 75%	6/8= 75%	As Total		As Total	As Total
225	90	21/28= 75%	10/28= 35.7%	4/8= 50%	3/5= 60%	1/2= 50%	3/6= 50%	As Total	As Total
226	174	19/24= 79.2%	10/24= 41.7%	4/7= 57.1%	3/3= 100%	None Known	As Total	As Total	As Total
227	127	23/26= 88.5%	16/26= 61.5%	1/5= 20%	1/2= 50%	None Supplied	As Total	As Total	As Total
228	53	18/18= 100%	17/18= 94.4%	3/10= 30%	3/8= 37.5%	1/5= 20%	2/5= 40%	As Total	As Total
229	61	10/29= 34.5%	5/29= 17.2%	1/4= 25%	1/4= 25%	1/1= 100%	1/4= 25%	As Total	As Total
230	23	7/17= 41.2%	4/17= 23.5%	1/4= 25%	1/3= 33.3%	0/1= 0	1/4= 25%	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
232	447	16/29= 55.2%	9/29= 31%	6/15= 40%	4/9= 44.4%	2/10= 20%	4/5= 80%	5/12= 41.7%	3/7= 42.9%
233	226	8/23= 34.8%	4/23= 17.4%	5/5= 100%	3/3= 100%	None Known	As Total	As Total	As Total
234	251	8/18= 44.4%	4/18= 22.2%	4/6= 66.7%	1/3= 33.3%	None Supplied	As Total	As Total	As Total
235	75	3/36= 8.3%	0/36= 0	4/6= 66.7%	3/4= 75%	1/1= 100%	3/5= 60%	As Total	As Total
236	344	11/17= 64.7%	3/17= 17.6%	8/8= 100%	8/8= 100%	None Supplied	As Total	As Total	As Total
237	500	8/25= 32%	3/25= 12%	6/6= 100%	6/6= 100%	None Supplied	As Total	As Total	As Total
238	87	18/23= 78.3%	6/23= 26.1%	6/7= 85.7%	5/6= 83.3%	None Supplied	As Total	As Total	As Total
239	264	10/15= 66.7%	8/15= 53.3%	8/8= 100%	8/8= 100%	1/1= 100%	7/7= 100%	As Total	As Total
240	121	6/20= 30%	5/20= 25%	4/4= 100%	4/4= 100%	None Known	As Total	As Total	As Total
242	304	5/25= 20%	2/25= 8%	12/13= 92.3%	9/9= 100%	5/5= 100%	7/8= 87.5%	As Total	As Total
243	1	1/1= 100%	1/1= 100%	0/1= 0	0/1= 0	None Known in Base	As Total	As Total	As Total
244	408	4/16= 25%	2/16= 12.5%	5/6= 83.3%	2/2= 100%	None Supplied	As Total	As Total	As Total
245	46	15/24= 62.5%	8/24= 33.3%	3/4= 75%	2/3= 66.7%	None Known	As Total	As Total	As Total
246	133	14/29= 48.3%	4/29= 13.8%	10/11= 90.9%	9/9= 100%	As Total		As Total	As Total
247	2	2/2= 100%	2/2= 100%	2/6= 33.3%	2/2= 100%	None Known	As Total	As Total	As Total
248	339	11/17= 64.7%	11/17= 64.7%	2/7= 28.6%	2/7= 28.6%	None Known	As Total	As Total	As Total
249	480	24/25= 96%	16/25= 64%	5/5= 100%	5/5= 100%	As Total		As Total	As Total
250	68 (80)	9/20= 45%	4/20= 20%	6/6= 100%	1/1= 100%	None Supplied	As Total	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
251	105	7/27= 25.9%	3/27= 11.1%	3/6= 50%	3/6= 50%	2/4= 50%	2/3= 66.7%	As Total	As Total
252	4	4/4= 100%	4/4= 100%	1/4= 25%	1/2= 50%	None Known	As Total	As Total	As Total
256	333	16/27= 59.3%	8/27= 29.6%	3/4= 75%	1/1= 100%	1/1= 100%	2/3= 66.7%	As Total	As Total
259	426	4/23= 17.4%	0/23= 0	11/12= 91.7%	2/2= 100%	8/9= 88.9%	3/3= 100%	4/4= 100%	2/2= 100%
260	51	12/23= 52.2%	6/23= 26.1%	3/5= 60%	2/4= 50%	None Known	As Total	As Total	As Total
263	14	4/11= 36.4%	2/11= 18.2%	1/1= 100%	1/1= 100%	None Supplied	As Total	As Total	As Total
264	278	18/26= 69.2%	6/26= 23.1%	3/8= 37.5%	3/6= 50%	None Known	As Total	As Total	As Total
266	29	9/15= 60%	5/15= 33.3%	2/3= 66.7%	1/1= 100%	1/2= 50%	1/1= 100%	As Total	As Total
268	10	1/10= 10%	1/10= 10%	1/3= 33.3%	1/3= 33.3%	None Known in Base	As Total	As Total	As Total
269	22	6/11= 54.5%	6/11= 54.5%	7/13= 53.8%	7/10= 70%	4/10= 40%	4/4= 100%	4/4= 100%	4/4= 100%
270	9	9/9= 100%	1/9= 11.1%	No Recall Base		None Known	None Found		
271	13 (163)	6/10= 60%	5/10= 50%	0/1= 0		None Supplied	As Total	As Total	
272	657	13/16= 81.2%	5/16= 31.2%	9/12= 75%	8/11= 72.7%	1/2= 50%	8/10= 80%	As Total	As Total
273	27	5/20= 25%	2/20= 10%	0/20= 0	0/16= 0	0/16= 0	0/5= 0	0/16= 0	0/13= 0
274	7	6/7= 85.7%	4/7= 57.1%	3/3= 100%	3/3= 100%	2/2= 100%	3/3= 100%	As Total	As Total
276	8	5/6= 83.3%	2/6= 33.3%	2/4= 50%	1/2= 50%	0/1= 0	2/3= 66.7%	As Total	As Total
277	500 (799)	21/23= 91.3%	11/23= 47.8%	16/16= 100%	15/15= 100%	As Total		As Total	As Total
278	91	16/21= 76.2%	6/21= 28.6%	3/10= 30%	3/9= 33.3%	1/7= 14.3%	2/3= 66.7%	3/4= 75%	3/4= 75%

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
281	343	14/20= 70%	4/20= 20%	7/9= 77.8%	5/7= 71.4%	1/1= 100%	6/8= 75%	As Total	As Total
301	489	3/18= 16.7%	2/18= 11.1%	14/27= 51.9%	12/19= 63.2%	10/17= 58.8%	6/13= 46.2%	10/19= 52.6%	9/15= 60%
302	2	0/2= 0	0/2= 0	0/0= 100%	0/0= 100%	None Known	None Found	As Total	As Total
303	36	0/22= 0	0/22= 0	0/2= 0		As Total		As Total	
304	56	10/18= 55.6%	6/18= 33.3%	6/8= 75%	5/6= 83.3%	5/6= 83.3%	3/4= 75%	3/4= 75%	3/3= 100%
305	405	19/26= 73.1%	15/26= 57.7%	1/8= 12.5%	1/4= 25%	As Total		As Total	As Total
306	93	11/25= 44%	3/25= 12%	2/6= 33.3%	1/1= 100%	0/1= 0	2/5= 40%	2/5= 40%	1/1= 100%
307	12	11/11= 100%	3/11= 27.3%	0/9= 0		0/1= 0	0/8= 0	As Total	
308	190	13/24= 54.2%	3/24= 12.5%	2/3= 66.7%		None Supplied	As Total	As Total	
451	50	4/22= 18.2%	0/22= 0	No Recall Base		None Known	None Found		
452	131	9/25= 36%	5/25= 20%	1/4= 25%	1/4= 25%	None Supplied	As Total	As Total	As Total
453	33	10/25= 40%	5/25= 20%	1/3= 33.3%		None Supplied	As Total	As Total	
454	71	2/27= 7.4%	1/27= 3.7%	0/1= 0		None Known	As Total	As Total	
455	173	2/19= 10.5%	1/19= 5.3%	1/6= 16.7%	1/5= 20%	1/3= 33.3%	0/3= 0	1/4= 25%	1/3= 33.3%
456	267	19/28= 67.9%	8/28= 28.6%	3/7= 42.9%	1/3= 33.3%	None Supplied	As Total	As Total	As Total
457	245	6/25= 24%	3/25= 12%	7/8= 87.5%	4/4= 100%	1/1= 100%	6/7= 85.7%	As Total	As Total
458	722	2/22= 9.1%	1/22= 4.5%	18/20= 90%	8/9= 88.9%	16/16= 100%	3/5= 60%	3/5= 60%	1/2= 50%
460	207	14/18= 77.8%	10/18= 55.6%	13/17= 76.5%	10/11= 90.9%	8/9= 88.9%	6/9= 66.7%	6/10= 60%	5/6= 83.3%

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
461	165	5/26= 19.2%	1/26= 3.8%	3/6= 50%	1/2= 50%	1/2= 50%	2/4= 50%	3/5= 60%	1/1= 100%
462	24	8/16= 50%	4/16= 25%	3/6= 50%	2/3= 66.7%	2/3= 66.7%	2/5= 40%	As Total	As Total
463	42	0/25= 0	0/25= 0	0/0= 100%	0/0= 100%	None Known	None Found	As Total	As Total
465	502	25/25= 100%	2/25= 8%	1/1= 100%	1/1= 100%	None Known in Base	As Total	As Total	As Total
466	133	20/22= 90.9%	18/22= 81.8%	9/9= 100%	9/9= 100%	1/1= 100%	8/8= 100%	As Total	As Total
467	51	2/19= 10.5%	0/19= 0	2/3= 66.7%	0/1= 0	2/2= 100%	0/1= 0	As Total	As Total
469	45	7/29= 24.1%	3/29= 10.3%	6/7= 85.7%	4/4= 100%	3/3= 100%	3/4= 75%	As Total	As Total
470	96	11/26= 42.3%	5/26= 19.2%	4/5= 80%	3/4= 75%	2/2= 100%	2/3= 66.7%	As Total	As Total
471	233	4/23= 17.4%	3/23= 13%	7/9= 77.8%	2/4= 50%	2/2= 100%	5/7= 71.4%	As Total	As Total
472	58	20/24= 83.3%	11/24= 45.8%	4/8= 50%	4/8= 50%	None Known in Base	As Total	As Total	As Total
473	110	6/25= 24%	2/25= 8%	6/6= 100%	1/1= 100%	None Known in Base	As Total	As Total	As Total
474	34	11/28= 39.3%	5/28= 17.9%	0/2= 0	0/2= 0	None Known	As Total	As Total	As Total
477	75	10/22= 45.5%	0/22= 0	8/8= 100%	5/5= 100%	4/4= 100%	5/5= 100%	5/5= 100%	3/3= 100%
478	211	15/25= 60%	7/25= 28%	3/3= 100%	3/3= 100%	As Total		As Total	As Total
479	69	3/24= 12.5%	1/24= 4.2%	0/5= 0	0/4= 0	As Total		As Total	As Total
480	101	13/18= 72.2%	9/18= 50%	4/5= 80%	3/3= 100%	0/1= 0	4/4= 100%	As Total	As Total
481	76	10/17= 58.8%	1/17= 5.9%	3/3= 100%	2/2= 100%	None Known	As Total	As Total	As Total
482	1	1/1= 100%	1/1= 100%	1/3= 33.3%	1/2= 50%	1/1= 100%	0/2= 0	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
483	266	7/12= 58.3%	5/12= 41.7%	7/13= 53.8%	7/12= 58.3%	4/8= 50%	3/5= 60%	3/5= 60%	3/4= 75%
484	309 (367)	17/22= 77.3%	12/22= 54.5%	8/8= 100%	8/8= 100%	None Supplied	As Total	As Total	As Total
485	19	5/18= 27.8%	2/18= 11.1%	0/4= 0	0/3= 0	None Known in Base	As Total	As Total	As Total
486	322	21/23= 91.3%	5/23= 21.7%	4/6= 66.7%	3/5= 60%	As Total		As Total	As Total
488	18	11/14= 78.6%	4/14= 28.6%	2/3= 66.7%	2/3= 66.7%	0/1= 0	2/2= 100%	As Total	As Total
489	5	0/4= 0	0/4= 0	0/0= 100%	0/0= 100%	None Known	None Found	As Total	As Total
490	370	9/24= 37.5%	2/24= 8.3%	4/4= 100%	1/1= 100%	1/1= 100%	3/3= 100%	As Total	As Total
491	12	2/12= 16.7%	0/12= 0	0/7= 0	0/6= 0	As Total		As Total	As Total
492	32	6/22= 27.3%	1/22= 4.5%	1/4= 25%	1/3= 33.3%	0/1= 0	1/3= 33.3%	As Total	As Total
493	20	13/13= 100%	13/13= 100%	0/11= 0	0/11= 0	As Total		As Total	As Total
495	96	12/30= 40%	9/30= 30%	8/12= 66.7%	7/10= 70%	As Total	6/7= 85.7%	As Total	A Total
496	22	15/20= 75%	7/20= 35%	1/2= 50%	1/1= 100%	None Known in Base	As Total	As Total	As Total
497	159	13/26= 50%	2/26= 7.7%	0/6= 0	0/4= 0	0/2= 0	0/4= 0	As Total	As Total
498	499	7/17= 41.2%	3/17= 17.6%	2/3= 66.7%	1/1= 100%	None Known in Base	As Total	As Total	As Total
499	33	13/22= 59.1%	5/22= 22.7%	3/4= 75%	2/2= 100%	None Known in Base	As Total	As Total	As Total
501	415	12/22= 54.5%	7/22= 31.8%	0/5= 0	0/4= 0	As Total		As Total	As Total
502	686	15/25= 60%	9/25= 36%	6/9= 66.7%	4/6= 66.7%	0/2= 0	6/7= 85.7%	As Total	As Total
503	434	1/26= 3.8%	0/26= 0	0/1= 0	0/1= 0	As Total	None Found	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
504	12	5/9= 55.6%	4/9= 44.4%	1/1= 100%	1/1= 100%	None Known in Base	As Total	As Total	As Total
505	70	16/23= 69.6%	9/23= 39.1%	8/14= 57.1%	5/8= 62.5%	6/9= 66.7%	2/5= 40%	4/8= 50%	3/6= 50%
506	25	14/20= 70%	9/20= 45%	3/9= 33.3%	2/7= 28.6%	2/4= 50%	1/5= 20%	1/5= 20%	0/3= 0
507	500 (857)	0/13= 0	0/13= 0	0/1= 0		None Known	0/1=0	As Total	
508	54	0/24= 0	0/24= 0	1/2= 50%		0/1= 0	1/1= 100%	As Total	
509	121	15/26= 57.7%	7/26= 26.9%	6/7= 85.7%	4/4= 100%	1/1= 100%	5/6= 83.3%	As Total	As Total
510	359	22/22= 100%	8/22= 36.4%	6/8= 75%	5/7= 71.4%	As Total		As Total	As Total
511	500 (503)	8/21= 38.1%	3/21= 14.3%	4/7= 57.1%	3/5= 60%	2/3= 66.7%	2/4= 50%	2/4= 50%	2/4= 50%
513	500	17/25= 68%	5/25= 20%	7/8= 87.5%	6/6= 100%	2/2= 100%	5/6= 83.3%	5/6= 83.3%	4/4= 100%
521	27	3/24= 12.5%	1/24= 4.2%	1/1= 100%	1/1= 100%	As Total	None Found	As Total	As Total
522	134	10/22= 45.5%	5/22= 22.7%	6/6= 100%	5/5= 100%	None Known	As Total	As Total	As Total
523	10	6/10= 60%	1/10= 10%	1/1= 100%	1/1= 100%	None Known	As Total	As Total	As Total
524	52	9/23= 39.1%	5/23= 21.7%	3/6= 50%	2/4= 50%	2/4= 50%	2/3= 66.7%	As Total	As Total
525	4	4/4= 100%	4/4= 100%	2/6= 33.3%	2/4= 50%	1/1= 100%	1/5= 20%	As Total	As Total
526	180	5/15= 33.3%	2/15= 13.3%	5/9= 55.6%		As Total		As Total	
527	231	11/25= 44%	7/25= 28%	2/9= 22.2%	2/7= 28.6%	0/3= 0	2/6= 33.3%	As Total	As Total
528	346	8/18= 44.4%	2/18= 11.1%	3/5= 60%	1/2= 50%	None Known	As Total	As Total	As Total
529	184	8/17= 47.1%	5/17= 29.4%	8/14= 57.1%	4/7= 57.1%	As Total		As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
530	10	3/8= 37.5%	3/8= 37.5%	4/11= 36.4%	4/8= 50%	4/5= 80%	1/7= 14.3%	1/7= 14.3%	1/4= 25%
531	43 (455)	8/16= 50%	3/16= 18.7%	0/1= 0		None Known	As Total	As Total	
532	15	9/10= 90%	3/10= 30%	2/5= 40%	1/3= 33.3%	0/1= 0	2/4= 50%	As Total	As Total
534	113	9/20= 45%	6/20= 30%	2/3= 66.7%	1/2= 50%	None Known	As Total	As Total	As Total
535	143	3/10= 30%	1/10= 10%	2/5= 40%	2/4= 50%	None Known	As Total	As Total	As Total
539	52	11/18= 61.1%	0/18= 0	8/11= 72.7%		As Total		2/3= 66.7%	
540	32 (135)	10/30= 33.3%	6/30= 20%	1/2= 50%	1/1= 100%	None Known	As Total	As Total	As Total
541	20	8/8= 100%	8/8= 100%	3/3= 100%	3/3= 100%	1/1= 100%	2/2= 100%	As Total	As Total
545	66	3/23= 13%	1/23= 4.3%	5/6= 83.3%	1/1= 100%	2/3= 66.7%	3/3= 100%	3/3= 100%	
547	417	9/22= 40.9%	4/22= 18.2%	6/9= 66.7%	5/8= 62.5%	As Total		As Total	As Total
548	58	15/30= 50%	11/30= 36.7%	3/4= 75%	3/4= 75%	1/1= 100%	2/3= 66.7%	As Total	As Total
551	21	12/21= 57.1%	10/21= 47.6%	5/6= 83.3%	4/4= 100%	1/1= 100%	4/5= 80%	As Total	As Total
553	9	6/9= 66.7%	5/9= 55.6%	2/3= 66.7%	2/2= 100%	2/2= 100%	0/1= 0	As Total	As Total
555	500 (897)	9/18= 50%	7/18= 38.9%	1/3= 33.3%	1/2= 50%	1/2= 50%	0/2= 0	As Total	As Total
557	108	2/30= 6.7%	1/30= 3.3%	3/5= 60%	2/2= 100%	1/3= 33.3%	2/2= 100%	As Total	As Total
559	0	0/0= 0	0/0= 0	0/3= 0		None Known	As Total	As Total	
560	19	8/17= 47.1%	4/17= 23.5%	1/4= 25%	1/4= 25%	0/2= 0	1/4= 25%	As Total	As Total
561	27	24/26= 92.3%	14/26= 53.8%	0/2= 0	0/1= 0	None Known	As Total	As Total	As Total

#	Total Retrieved	PRECISION RATIO		RECALL RATIO TOTAL		RECALL RATIO BREAKDOWN			
		Overall	Major Value	Overall	Major Value	Known by Requester	Found Manually	"Best Set" Overall	"Best Set" Major Value
567	500 (526)	21/25= 84%	2/25= 8%	4/8= 50%	1/3= 33.3%	0/2= 0	4/6= 66.7%	4/6= 66.7%	1/1= 100%
568	103	18/22= 81.8%	13/22= 59.1%	5/7= 71.4%	5/7= 71.4%	2/2= 100%	4/6= 66.7%	4/6= 66.7%	4/6= 66.7%
569	82	8/17= 47.1%	4/17= 23.5%	4/10= 40%	3/7= 42.9%	2/4= 50%	2/6= 33.3%	As Total	As Total
570	333	20/28= 71.4%	2/28= 7.1%	2/7= 28.6%	1/3= 33.3%	1/2= 50%	1/5= 20%	1/5= 20%	0/1= 0
603	121	0/19= 0	0/19= 0	1/3= 33.3%	1/2= 50%	None Supplied	As Total	As Total	As Total
606	336	24/26= 92.3%	11/26= 42.3%	1/9= 11.1%	1/5= 20%	As Total		As Total	As Total

Appendix 5

Specificity of search formulations:

(a) Formulations nonspecific (precision failures)

(b) Specific formulations (recall failures)

307-006 O-68—17

APPENDIX 5 (a)

Search formulations nonspecific in relation to stated requests

(100 searches examined. 27 found to include instances of precision failures due to nonspecific formulations).

	Actual broadened search		Search strictly matching request	
	Of value	Of no value	Of value	Of no value
#117	11	11	11	10
#118	5	21	5	12
#155	13	10	13	8
#166	7	2	7	1
#173	8	7	5	6
#174	10	10	6	1
#177	3	20	1	18
#182	9	14	0	3
#200	10	11	10	9
#212	4	8	4	3
#213	0	1	0	0
#216	6	22	2	20
#221	3	24	3	18
#224	14	8	14	7
#230	7	10	6	7
#236	11	6	2	1
#238	18	5	5	4
#240	6	14	4	10
#245	15	9	11	8
#260	12	11	11	9

	Actual broadened search		Search strictly matching request	
	Of value	Of no value	Of value	Of no value
#268	1	9	1	0
#301	3	15	2	0
#304	10	8	10	7
#462	8	8	7	4
#478	15	10	14	8
#479	3	17	3	11
#483	7	5	7	4
TOTALS	122	169	101	102

Failure to broaden the searching strategy would have avoided 67/169 (39.6%) of the irrelevant articles while missing 21/122 (17.2%) of the relevant items.

Search-by-search analysis

#117

Request Schizophrenia and autism: epidemiology, genetics, etiology.

Elaboration The searcher included FACTOR ANALYSIS, STATISTICAL, which is not specific to epidemiology (i.e., she substituted one technic term for another).

Results Recall was unaffected. Without the elaboration, retrieval of irrelevant documents would have been reduced by 1/11.

#118

Request Immunochemistry of polypeptides

Elaboration The requester appears interested in the immunochemical structure and antigenic structure of proteins and polypeptides. Several term combinations used by the searcher are nonspecific in relation to this request. The term TRYPSIN, presumably included as an enzyme that may act on peptide linkages, retrieved only irrelevant material. This is an example of an "agent" term substituted for the substance (peptide or protein) upon which it might act.

Also, the terms ANTIBODY FORMATION and ANTIGEN-ANTIBODY REACTIONS are not specific in relation to "antigenic structure". Combined with protein terms, they retrieved a number of irrelevant articles on immunization or studies of a clinical nature on human serum. This is an example of "process" terms substituted for "structure" terms.

Results Recall was not improved. Without the elaboration, retrieval of irrelevant documents would have been reduced by 9/21.

#155

Request Etiology of decompression sickness and nitrogen narcosis, including bubble nucleation, growth and stabilization, gas solubility and diffusivity in liquids and body tissue, and models of gas or solute exchange in the body.

Elaboration The searcher substituted a general property term (CHEMISTRY, PHYSICAL) for the specific properties (solubility and diffusivity) requested.

Results Recall was not improved. Without the elaboration, retrieval of irrelevant documents would have been reduced by 2/10.

#166

Request Hemangioma of the small intestine

Elaboration The disease category was elaborated to include HEMANGIOPERICYTOMA and HEMANGIOENDOTHELIOMA as well as HEMANGIOMA.

Results Recall unaffected. Without the elaboration, retrieval of irrelevant documents would have been reduced by 1/2.

#173

Request Correlation between dentistry and arthritis.

Elaboration The requester is interested in oral manifestations of arthritis, infectious arthritis, rheumatoid arthritis, rheumatic fever, osteoarthritis, and Reiter's disease. The disease category was expanded to include related syndromes (e.g., Sjogren's), gout, and spondylitis.

Results Without the elaboration, retrieval of irrelevant documents would have been reduced by 1/7 but retrieval of relevant documents would have been reduced by 3/8.

#174

Request Testicular biopsies in infertility and endocrine disease.

Elaboration The searcher exploded on two facets of the request simultaneously. This is likely to result in high recall but low precision. For the anatomical term TESTIS, disease terms (e.g., TESTICULAR DISEASES, TESTICULAR NEOPLASMS) and "component part" terms (e.g., LEYDIG CELLS, SERTOLI CELLS) were substituted. At the same time, there was a generalization in the technics facet from BIOPSY to general terms such as PATHOLOGY and HISTOLOGY.

Results Without the elaboration, retrieval of irrelevant documents would have been reduced by 9/10 while retrieval of relevant documents would have been reduced by 4/10.

#177

Request Obstetrical management of premature rupture of the fetal membranes.

Elaboration The requester is interested in a specific complication of pregnancy or labor, namely premature rupture of the fetal membranes. The searcher generalized in the "disease" category to "complications involving the fetal membranes", by the coordination of FETAL MEMBRANES and PREGNANCY COMPLICATIONS

or

LABOR COMPLICATIONS

Results Without the elaboration, retrieval of irrelevant documents would have been reduced by 2/20 while retrieval of relevant documents would have been reduced by 2/3.

#182

Request Life span (turnover time) of cells in the kidney.

Elaboration In this case the searcher was forced to be nonspecific because the precise term that equates with "life span", LONGEVITY, does not appear to have been applied to articles on cell lifetime. For this "dimensional" term the searcher substituted the "effect" terms AGING and NECROSIS and terms (TRITIUM and CARBON ISOTOPES) representing specific agents that may be used in the measurement of cell lifetime.

Results Without the elaboration, retrieval of relevant documents would have been reduced by 11/14. However, recall would have been zero.

#200

Request Endometrium: ultrastructure, fine structure, electron microscopy, histochemistry, cytochemistry.

Elaboration It is difficult to conduct a search on ultrastructure because there is no specific term denoting "fine structure" in the vocabulary. For terms indicating "structure" the searcher substituted terms representing "parts" by exploding on the term CYTOPLASM. Specificity was further reduced by substituting a "disease" term (ENDOMETRIAL HYPERPLASIA) for an organ term (ENDOMETRIUM).

Results This elaboration did not improve recall. Without the elaboration, retrieval of irrelevant documents would have been reduced by 2/11.

#212

Request Adherence of patients to oral tuberculosis therapy (isoniazid and PAS) as indicated by periodic urine testing.

Elaboration The requester is interested only in the specific metabolic processes of absorption and excretion. The searcher generalized to METABOLISM in general, and also substituted the entire FLUIDS and SECRETIONS group of terms (including, for example, SPUTUM) for the specific fluid, URINE, of interest.

Results Without the elaboration, retrieval of irrelevant documents would have been reduced by 5/8, while recall of relevant documents would not have been affected.

#213

Request Effect of certain diuretic agents (thiazides, ethacrynic acid, furosemide, and MK-870) on: (a) renal excretion of calcium, magnesium and phosphate, (b) gastrointestinal absorption of calcium, magnesium and phosphate, (c) metabolic bone changes on calcium, magnesium and phosphate.

Elaboration Inexplicably the search was conducted only on DIURETICS, ETHACRYNIC ACID and FUROSEMIDE, omitting many specific thiazide terms of interest to the requester. Only one citation was retrieved, on the generic term DIURETICS, and this proved to deal with malachite green, a substance of no relevance to the request.

Results Retrieval would have been zero without the inclusion of DIURETICS. The only article retrieved, however, was irrelevant (i.e., zero precision).

#216

Request Histology and ultrastructure of the olfactory mucosa.

Elaboration As noted previously (see analysis #200 above) it is difficult to conduct a successful search relating to anatomical "structure" because the vocabulary is deficient in terms in this area, although this situation became less acute with the introduction of the subheadings CYTOLOGY and ANATOMY & HISTOLOGY. Whereas in search #200, the analyst substituted "part" terms by exploding on CYTOPLASM, the formulation for this search relied mainly on "technic" terms by means of explosions on MICROSCOPY and HISTOLOGICAL TECHNICS. These explosions inevitably retrieved unwanted articles on methodology rather than findings. The searcher, unaccountably, made no use of subheadings, although four of the six articles known to be relevant, and indexed since the introduction of subheadings, could have been retrieved on NASAL MUCOSA/CYTOLOGY and NASAL MUCOSA/INNERVATION.

Results Without the elaboration, retrieval of irrelevant articles would have been reduced by 2/22 while retrieval of relevant articles would have been reduced by 4/6.

#221

Request Medical articles relating to the Somali republic, especially tropical medicine, parasitology, microbiology, mycology, and enteric infections.

Elaboration The searcher exploded on AFRICA, EASTERN (i.e., a movement from a specific geographical heading to a generic). While the move from SOMALIA to the single term AFRICA, EASTERN may be justified, it is difficult to condone an explosion on the latter term, which retrieves articles on other countries (e.g., ETHIOPIA) that can have no direct bearing on the subject of the search.

Results Without the elaboration, retrieval of irrelevant documents would have been reduced by 6/24. Recall would have been reduced by omitting AFRICA, EASTERN, but the additional geographical terms brought in by the explosion produced only noise.

#224

Request Nonsurgical diagnosis and therapy of primary bone tumors, benign and malignant.

Elaboration The searcher generalized in the "technic" category from "drug therapy" to an explosion on DRUG ADMINISTRATION. This covers terms (e.g., INJECTIONS, INTRAVENOUS) not exclusively diagnostic or therapeutic and retrieves articles on, for example, injection of carcinogenic agents into laboratory animals.

Results No improvement in recall. Without the elaboration, retrieval of irrelevant items would have been reduced by 1/8.

#230

Request Frozen serum and plasma.

Elaboration The searcher expanded in the "anatomical term" category by including THROMBOPLASTIN, BLOOD COAGULATION FACTORS (including specific factor terms) and other terms not strictly related to plasma or serum.

Results Without the elaboration, retrieval of irrelevant articles would have been reduced by 3/10 while retrieval of relevant articles would have been reduced by 1/7.

#236

Request Corneal preservation methods

Elaboration In addition to "preservation" terms, the searcher included "transplantation" terms in the formulation. That is, one group of technic terms (representing the use to which the preserved tissue is to be put) is substituted for another. Although some irrelevancies (articles dealing exclusively with technics of transplantation) will result, it seems essential to high recall that any search on preservation be expanded to cover transplantation.

Results Without the elaboration, retrieval of irrelevant articles would have been reduced by 5/6, while retrieval of relevant articles would have been reduced by 9/11.

#238

Request Heart preservation methods

Elaboration As above (#236). "Transplantation" terms were substituted for "preservation" terms.

Results Without the elaboration, retrieval of irrelevant articles would have been reduced by 1/5, while retrieval of relevant items would have been reduced by 13/18.

#240

Request Radiation kidney (radiation nephritis), human and experimental.

Elaboration The request asks for a specific renal effect of radiation, namely nephritis. The searcher generalized from the specific disease term indicated by coordinating radiation terms with anatomic terms (kidney and kidney parts) and with a technic term (KIDNEY FUNCTION TESTS).

Results Without this elaboration, retrieval of irrelevant articles would have been less by 4/14, while retrieval of relevant articles would have been reduced by 2/6.

#245

Request Radiation pneumonitis and radiation pulmonary fibrosis, including pathology, physiology, radiography and therapy.

Elaboration As for #240, the searcher substituted anatomical terms (an explosion on LUNG) for the specific disease terms implied by the request. That is, she generalized from "radiation pneumonitis and radiation pulmonary fibrosis" to "adverse effects of radiation on the lung".

Results Without the elaboration, retrieval of irrelevant items would have been reduced by 1/9, while retrieval of relevant items would have fallen by 4/15.

#260

Request Reticuloendotheliosis (microgliomatosis, reticulosarcoma) of the retina or brain.

Elaboration By coordinating SARCOMA, RETICULUM CELL or RETICULOENDOTHELIOSIS with EYE NEOPLASMS, the searcher is substituting a disease term for an anatomical term (RETINA) and thus expanding the search from "reticuloendotheliosis of the retina" to "reticuloendotheliosis of the eye".

Results Without this generalization, retrieval of irrelevant items would have been less by 2/11, while retrieval of relevant items would have fallen by 1/10.

#268

Request Aspergillosis of the orbit

Elaboration The area of interest relates to one particular mycosis (covered by the term ASPERGILLOSIS) caused by one particular fungus (covered by the term ASPERGILLUS). It is therefore difficult to justify the searcher's expansion to "all fungal diseases of the orbit" by explosions on MYCOSES and FUNGI.

Results These explosions added nothing to the recall but sharply reduced precision from 1/1 (100%) to 1/10 (10%).

#301

Request Methods of evaluating performance of surgical sutures in vivo, methods of introducing infection into sutured wounds, and adverse effects of various suture materials.

Elaboration The searcher expanded the scope of the search from "adverse effects of suture materials" to "adverse effects of suture materials and suturing technics" -- by including the terms SUTURE TECHNICS and LIGATION. That is, "technic" terms are substituted for materials terms.

Results By omission of the elaboration, retrieval of irrelevant documents would have been reduced 100% (15/15), while retrieval of relevant documents would have dropped by 1/3.

#304

Request Effect of administration of chloramphenicol on pregnant females during the first trimester: animal or human studies.

Elaboration The requester is concerned with congenital abnormalities attributable to chloramphenicol. This was generalized to "chloramphenicol-induced abnormalities" (not necessarily congenital) by the coordination of CHLORAMPHENICOL and ABNORMALITIES.

Results Recall was not improved. Without the elaboration, retrieval of irrelevant articles would have been reduced by 1/8.

#462

Request Cerebral amyloidosis

Elaboration As well as coordinating AMYLOIDOSIS or AMYLOID SUBSTANCE with anatomical terms relating to the brain, the searcher also coordinated with "brain disease" terms. This brought out additional relevant articles, but also retrieved irrelevant items; for example, renal amyloidosis in paraplegic patients.

Results Without the elaboration, retrieval of irrelevant articles would have been reduced by 4/8, while retrieval of relevant articles would have fallen by 1/8.

#478

Results Experimental wound healing of skin, subcutaneous tissue, intestine and gingiva.

Elaboration In an attempt to cover the aspect of wound healing in skin transplantation, the requester used a coordination of SKIN TRANSPLANTATION and WOUND HEALING or REGENERATION. That is, from the hierarchical tree G1.92:

REGENERATION
BONE REGENERATION
LIVER REGENERATION
NERVE REGENERATION
WOUND HEALING

the generic REGENERATION was substituted for the specific WOUND HEALING. REGENERATION and SKIN TRANSPLANTATION retrieved irrelevant articles, largely immunological studies, on graft rejection. The requester is not concerned with immunological studies nor indeed with skin transplantation per se unless the articles contain a substantial amount of information on the mechanism of wound healing.

Results Without the expansion, retrieval of irrelevant articles would have been reduced by 2/10, while retrieval of relevant articles would have been reduced by 1/15.

#479

Request Quantitative and kinetic aspects of viral antigen-antibody reactions.

Elaboration To be specific, the searcher needed to specify that an antigen or antibody term should co-occur with a virus term and also with a term indicating kinetics or quantitative analysis. However, in some sub-searches a testing "technic" term was substituted for an antibody ("substance") term, in particular by exploding on SERODIAGNOSIS. This explosion brings in terms not exclusively related to antigen-antibody reactions. The combination RUBELLA and NEUTRALIZATION TESTS and MODELS, for example, retrieved irrelevant articles on such topics as experimental rubella virus infections in rhesus monkeys.

Results Without the elaboration, retrieval of irrelevant documents would have been reduced by 6/17, while retrieval of relevant documents would not have been affected.

#483

Request Organ preservation, storage and banking.

Elaboration The request is for "organ preservation" and not for "tissue preservation". The searcher departed from organs in the strict sense (the sense in which the term was used by the requester) by including such terms as BONE MARROW.

Results Recall unaffected. Without the elaboration retrieval of irrelevant documents would have been reduced by 1/5.

APPENDIX 5 (b)

Specificity of search formulation

In a sample of 210 searches analyzed, only 9 were found in which recall failures were attributed directly to the specificity of a search formulation. In no case was the formulation more specific than warranted by the stated request. Rather, these are examples of (1) situations in which the searcher might reasonably have been expected to generalize in order to improve recall or (2) situations in which search generalization would be necessary in order to substantially improve recall. A brief summary of the analysis is given below.

38 Hospital medical staff organization and functions, including the infection committee, the quality of medical care, and continuing education in the hospital.

This search was conducted on the single terms HOSPITAL MEDICAL STAFF and EDUCATION, MEDICAL, CONTINUING. This retrieved a total of 492 items, of which approximately 65% were of some relevance. Because of the paucity of terms in the area of hospital management, and because hospital journals tend not to be indexed exhaustively, the searcher would have needed to search very broadly on HOSPITAL ADMINISTRATION (over 1300 postings), in addition to the two terms used, in order to approach 100% recall. Precision would, of course, have been very low.

48 Rehabilitation of hemiplegics; psychology of hemiplegia patients.

We have a low recall estimate of 38.5% for this search, largely because much of the relevant literature deals with rehabilitation of the "stroke" patient in general, rather than the hemiplegic in particular. The addition of the terms CEREBRAL HEMORRHAGE and CEREBROVASCULAR DISORDERS, as alternates to HEMIPLEGIA, would have raised recall to over 60%. Such a generalization might have improved recall to almost 80% if the term REHABILITATION had been applied to a number of articles to which it appears pertinent.

77 Stomach neoplasms: epidemiology, etiology, genetics

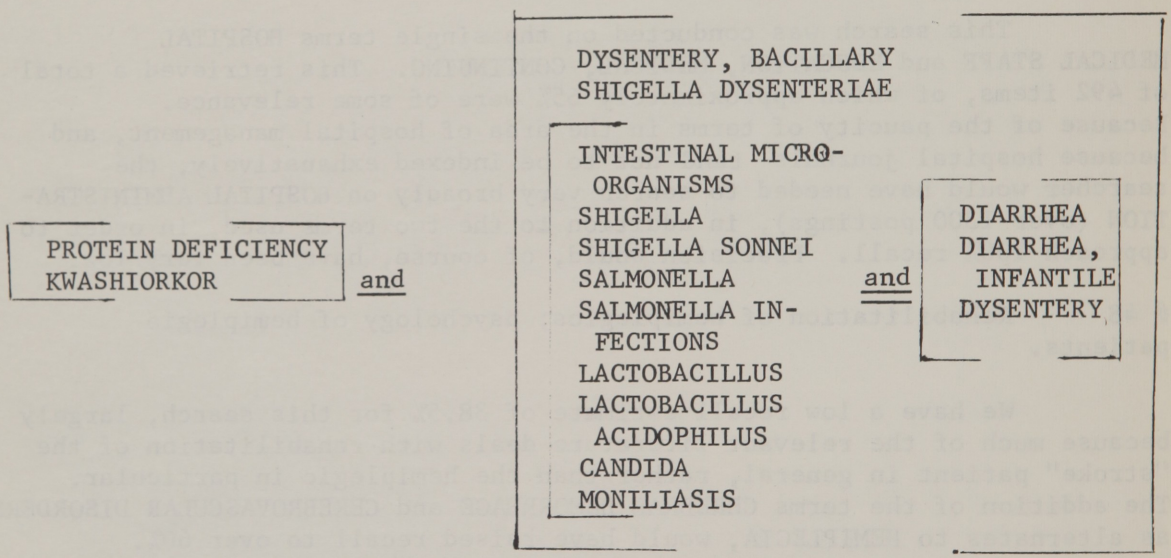
The searcher used only the terms STOMACH NEOPLASMS and CARCINOMA, SCIRRHOUS in the disease category, but coordinated these terms with a long list of terms in an attempt to cover all approaches to epidemiology and etiology. Consequently, there were many term combinations that were not specific to the request (e.g., STOMACH NEOPLASMS and NEGROES, STOMACH NEOPLASMS and MORTALITY), and precision was only 44% (294 articles retrieved). In this case, the searcher might have tried broadening in the disease category to GASTROINTESTINAL NEOPLASMS, coordinating this term with others strictly related to epidemiology or etiology. Recall might thus have been improved from 66% to around 73%.

82 Factor analysis of surgery of temporal lobe epilepsy. Specific factors are related to radiography, EEG, psychological tests used before and after surgery, and prognosis.

The searcher coordinated all epilepsy terms with various surgery terms, but only the terms EPILEPSY, TEMPORAL LOBE, and EPILEPSY, TRAUMATIC with the appropriate diagnostic terms. It would be necessary to generalize to EPILEPSY, as a coordinate with the diagnostic terms, in order to improve recall substantially (from 65% to 85% approximately). The total number of documents retrieved would, of course, be much greater than the 315 actually retrieved, and the precision ratio would probably be much lower than the 37% actually attained in this search.

115 Shigella, salmonella, lactobacillus, candida, or intestinal microorganisms causing diarrhea or dysentery in cases of protein deficiency or Kwashiorkor.

The searcher relied on a specific and exhaustive formulation, as follows:



This retrieved nothing although there are articles of varying degrees of relevance within MEDLARS. It would be necessary to generalize to such combinations as KWASHIORKOR and GASTROENTERITIS in order to obtain high recall in this search.

152 Anticoagulants for the treatment or prevention of pulmonary embolus.

The search was conducted on PULMONARY EMBOLISM and a list of anticoagulant terms. Recall would have improved, but only marginally (from 27% to about 36%), by generalizing to THROMBOEMBOLISM.

155 Etiology of decompression sickness and nitrogen narcosis, including bubble nucleation, growth and stabilization, gas solubility and diffusivity, and analogue models or inert gas exchange.

The searcher might reasonably have generalized from individual inert gas terms to GASES in view of the fact that there is no intermediate generic term for "inert gases" as a group. However, this would only have improved recall from 50% to about 58%.

166 Hemangioma of the small intestine

It appears that there are very few articles directly on the subject of hemangioma of the small bowel, although there are general articles on small bowel neoplasms that devote some attention to hemangioma, although not necessarily indexed under this specific term. To retrieve these the searcher might well have expanded to INTESTINAL NEOPLASMS and INTESTINE, SMALL.

#203 The use of gamma globulin in the treatment of bacteriologic diseases.

The searcher coordinated gamma globulin terms with terms for individual bacteriologic diseases. Unfortunately, some diseases (e.g., ASTHMA) may or may not be bacterial in origin. However, the index language does not distinguish the bacterial version from a version of different etiology. It would be necessary to expand to the whole of infectious disease, coordinated with gamma globulin terms, in order to obtain reasonable recall in this search.

Appendix 6

Subheadings: application to cases of false coordination and incorrect term relationship

Appendix 6

Analysis of failures due to false coordinations and incorrect term relationships: possibility of solution by use of subheadings.

Number of searches analyzed: 45

False coordinations

Number of distinct examples encountered in 45 searches: 20

Number solvable by use of existing subheadings: 12

Number solvable by use of suggested new subheadings: 4

Number not readily solvable by subheadings: 4

Incorrect term relationships

Number of distinct examples encountered in 45 searches: 22

Number solvable by use of existing subheadings: 14

Number solvable by use of suggested new subheadings: 6

Number not readily solvable by subheadings: 2

307-006 O-68—18

#230

Request Blood preservation: frozen serum and plasma

Formulation

BLOOD PRESERVATION and GLYCERIN or FREEZING and SERUM ALBUMIN or PLASMA

Unwanted articles The "preservation" terms do not relate to the serum or plasma terms (e.g., articles dealing with preserved red blood cells transformed to plasma).

Solution A new subheading PRESERVATION applicable to all tissue terms in MeSH.

#239

Request Experimental renal hypertension in any animal

Formulation HYPERTENSION, RENAL and RAT

Unwanted articles There is no direct relation between "hypertension" and "rat". Determination of angiotension II in rat blood and the significance of this in the study of human hypertension.

Solution A new subheading EXPERIMENTAL. When added to a disease term, it would tend to ensure that this term is directly related to an animal term applied in indexing. HYPERTENSION, RENAL/EXPERIMENTAL and RAT would have avoided the false coordination mentioned above.

#240

Request Radiation damage to the kidney (radiation nephritis)

Formulation RADIATION INJURY and HYDRONEPHROSIS or KIDNEY DISEASES

Unwanted articles The "radiation injury" and kidney disease terms refer to two different patients.

Solution The existing subheading ETIOLOGY, applied to the kidney disease terms, would tend to reduce false coordinations of this type. A new subheading RADIATION-INDUCED would eliminate them.

#260

Request Reticuloendotheliosis of retina or brain

Formulation SARCOMA, RETICULUM CELL and OPHTHALMIC NERVE

Unwanted articles Neurological complications of malignant tumors of the nasopharynx. The two terms refer to different patients.

Solution There seems no simple way to avoid this type of false coordination by the conventional use of subheadings.

#304

Request Effect of administration of chloramphenicol to pregnant female during first trimester.

Formulation CHLORAMPHENICOL and PREGNANCY or ABNORMALITIES or ABORTION, SEPTIC

Unwanted articles Two types: 1. the chloramphenicol is not directly related to pregnancy; it is used therapeutically after septic abortion or after surgery for kidney abnormalities; 2. article on glucose-6-phosphate-dehydrogenase deficiency in Thailand; fertility, neonatal jaundice, bilirubin encephalopathy and acute hemolytic anemia are discussed, chloramphenicol being mentioned as a possible etiological factor in the latter (there is no reference to chloramphenicol in pregnancy).

Solution Use of CHLORAMPHENICOL/ADVERSE EFFECTS would have avoided retrieval of the first group, in which the relationship is that of CHLORAMPHENICOL/THERAPEUTIC USE only. This subheading could not, however, prevent the second type of false coordination without joint use of the subheading DRUG EFFECTS. CHLORAMPHENICOL/ADVERSE EFFECTS and PREGNANCY/DRUG EFFECTS would have successfully avoided the second type of false coordination.

#452

Request Complications (i.e., sequelae) of pancreatitis.

Formulation PANCREATITIS and CHOLELITHIASIS or JAUNDICE, OBSTRUCTIVE or COMMON BILE DUCT CALCULI

Unwanted articles Various diagnostic procedures in hepatobiliary disease. In each article, some patients have pancreatitis, others have other conditions. There is no coincidence of pancreatitis and jaundice or calculi or gallstones in the same patient.

Solution The existing subheading COMPLICATIONS can solve this type

of problem when applied to both disease terms anded in the formulation. Thus, the use of PANCREATITIS/COMPLICATIONS and CHOLELITHIASIS/COMPLICATIONS, in a search formulation, will tend to retrieve articles in which the two terms relate to the same patient. It seems preferable, however, to use the subheading COMPLICATIONS solely for the concept of co-existing diseases, and to introduce a new subheading SEQUELAE to be used exclusively for conditions following as a consequence of other diseases and, possibly, conditions following as a consequence of various forms of therapy, exercise, etc.

#453

Request Perforation of the gallbladder

Formulation CHOLELITHIASIS and RUPTURE, SPONTANEOUS
or
PERITONITIS

Unwanted articles Surgical and diagnostic studies in which the two terms are not directly related (i.e., the rupture or peritonitis does not follow as a consequence of the gallstones). For example, in one article the term RUPTURE, SPONTANEOUS applies to blood vessel rupture.

Solution In this case the searcher is interested in rupture or peritonitis solely as sequelae of cholelithiasis. A subheading SEQUELAE, appended to CHOLELITHIASIS, would have avoided this type of false coordination. The subheading ETIOLOGY may be used as a companion subheading in this context. That is, to express condition B following as a consequence of condition A, we can use A/SEQUELAE and B/ETIOLOGY (or A/COMPLICATIONS and B/ETIOLOGY).

#458

Request Effect of exercise on respiratory, circulatory and cardiovascular systems.

Formulation EXERTION and HYPERVENTILATION

Unwanted articles Discussion on gastrectomy: hyperventilation is used in assessing the pulmonary status of surgical patients; exercise is used in postoperative therapy. That is, there is no reference to hyperventilation following exertion.

Solution EXERTION/SEQUELAE and EXERTION/ADVERSE EFFECTS are the only relationships relevant to the request. They would prevent the false coordination mentioned above, in which the relationship is that of EXERTION/THERAPEUTIC USE (the term EXERCISE THERAPY is correct here).

#462

Request Cerebral amyloidosis.

Formulation AMYLOIDOSIS and CEREBRAL SCLEROSIS, DIFFUSE

Unwanted articles Rectal biopsy: no direct relation between the two terms.

Solution There appears to be no simple way to avoid this type of false coordination by the conventional use of subheadings. On the other hand, the coordination of AMYLOIDOSIS with a brain disease term, to express "cerebral amyloidosis", is a searching strategy of doubtful value.

#457

Request Neurological complications of kidney diseases.

Formulation UREMIA and MYOSITIS

Unwanted articles Serum calcium measurements in various diseases, including uremia and polymyositis (the two conditions do not refer to the same patient).

Solution UREMIA/COMPLICATIONS and MYOSITIS/COMPLICATIONS

or

UREMIA/COMPLICATIONS and MYOSITIS/ETIOLOGY ("SEQUELAE" would be preferable).

#466

Request Lung transplantation

Formulation LUNG and TRANSPLANTATION

Unwanted articles Distribution of colloidal gold in rats with transplanted Guerin tumor (i.e., no lung transplantation).

Solution Use of existing subheading TRANSPLANTATION. LUNG/TRANSPLANTATION would have avoided this false coordination.

#473

Request Neurological and muscular complications of chickenpox or varicella. Infections precipitating myasthenia gravis.

Formulation MYASTHENIA GRAVIS and infection terms

CHICKEN POX
VARICELLA-ZOSTER VIRUS and neurological disease terms

Unwanted articles The terms are not related (i.e., do not refer to the same patient).

Solution

CHICKEN POX/SEQUELAE and neurological disease term/ETIOLOGY

infection term/SEQUELAE and MYASTHENIA GRAVIS/ETIOLOGY

A problem still remains. Coordination of a virus term and MYASTHENIA GRAVIS retrieves articles in which the two terms are essentially unrelated. Assuming that a subheading SEQUELAE would be out of place appended to a virus term, the combination MYASTHENIA GRAVIS/ETIOLOGY and a virus term would still tend to ensure that the virus is an etiological factor in the disease. The existing subheading COMPLICATIONS can be used in place of SEQUELAE but is less precise.

#474

Request Intestinal atony or ileus following vagotomy and pyloroplasty.

Formulation PYLORIC STENOSIS and SURGERY, OPERATIVE
POSTOPERATIVE COMPLICATIONS

and INTESTINAL OBSTRUCTION. OR

VAGOTOMY and INTESTINAL OBSTRUCTION

Unwanted articles Case series in which the pyloric stenosis or vagotomy and the intestinal obstruction refer to different patients.

Solution 1. VAGOTOMY/ADVERSE EFFECTS and INTESTINAL
OBSTRUCTION/ETIOLOGY

2. PYLORIC STENOSIS/SURGERY and SURGERY, OPERATIVE/ADVERSE EFFECTS

and INTESTINAL OBSTRUCTION/ETIOLOGY

This latter formulation is rather cumbersome. The problem could better be solved by means of a new subheading POSTOPERATIVE COMPLICATIONS. PYLORIC STENOSIS/POSTOPERATIVE COMPLICATIONS and INTESTINAL OBSTRUCTION/ETIOLOGY would tend to prevent the false coordinations noted above.

#478

Request Experimental wound healing of skin, gingiva, or intestine.

Formulation DUODENUM and WOUND HEALING

Unwanted articles Not duodenal wound healing but healing of the Ampulla of Vater after sphincterotomy, with and without suture of the duodenal mucosa.

Solution DUODENUM/INJURIES and WOUND HEALING would prevent this type of false coordination.

#480

Request Instantaneous readoff of diagnosis or pattern of electrocardiograms by computer.

Formulation AUTOMATIC DATA PROCESSING and ELECTROCARDIOGRAPHY

Unwanted articles Electronics in gynecology and obstetrics: fetal heart recording and ADP of obstetrical records.

Solution ELECTROCARDIOGRAPHY/INSTRUMENTATION and
AUTOMATIC DATA PROCESSING

#484

Request Complications following the use of oral contraceptives.

Formulation MEDROXYPROGESTERONE and STERILITY, FEMALE

Unwanted articles Use of provera to delay implantation in rats. Phenylethylamines were tested to determine their antifertility effects.

Solution MEDROXYPROGESTERONE/ADVERSE EFFECTS

#489

Request Adreno-hepatic fusion (i.e., adrenal gland fused with the overlying liver)

Formulation ADRENAL GLANDS/ABNORMALITIES and LIVER/ABNORMALITIES

Unwanted articles Liver abnormalities and adrenal abnormalities discussed separately.

Solution There appears no simple way of avoiding this by the conventional use of subheadings. This type of false coordination will occur occasionally but should be within tolerable limits. Note: this is the only example discovered in which a false coordination occurred despite the correct use of subheadings in indexing and searching.

#513

Request Effect of irradiation on the infectivity and structure of viruses.

Formulation ONCOGENIC VIRUSES and RADIATION GENETICS

HERPESVIRUS INFECTIONS and ULTRAVIOLET RAYS

Unwanted articles The former term combination retrieved an article

on cancer etiology in which radiation and viruses are not directly related. The latter discusses a vaccine, inactivated by ultraviolet, used to protect against herpetic virus.

Solution 1. ONCOGENIC VIRUSES/RADIATION EFFECTS

2. The combination HERPESVIRUS INFECTIONS and ULTRAVIOLET RAYS was presumably included in the search formulation to retrieve articles relating to effect of ultraviolet on infectivity by herpesvirus, where the indexer has used the term HERPESVIRUS INFECTIONS and ULTRAVIOLET RAYS. There seems no good way of preventing a false coordination here (unless, of course, the subheading RADIATION EFFECTS is made applicable to infection terms, but this would conflict with the heading RADIOTHERAPY). Reliance on HERPES VIRUS/RADIATION EFFECTS would have prevented this failure.

INCORRECT TERM RELATIONSHIPS

#240

Request "Radiation kidney"

Formulation RADIATION INJURY and NEPHROBLASTOMA

Unwanted articles Not radiation-induced nephroblastoma but therapeutic use of radiation following nephrectomy.

Solution RADIATION INJURY/SEQUELAE and NEPHROBLASTOMA/ETIOLOGY

#249

Request Primary tumors or cysts of the heart, pericardium, and major blood vessels.

Formulation NEOPLASM RADIOTHERAPY and VENA CAVA, SUPERIOR

Unwanted articles Not radiotherapy of neoplasms of vena cava, but vena caval syndrome resulting from neoplasms of the lung.

Solution HEART NEOPLASMS/RADIOTHERAPY

#260

Request Reticuloendotheliosis of the retina or brain

Formulation SARCOMA, RETICULUM CELL and BRAIN DISEASES

Unwanted articles Reticulum cell sarcoma, with encephalopathy.

Solution There seems no convenient way of preventing this failure by use of subheadings. However, the coordination of two disease terms to express "reticuloendotheliosis of the brain" is a strategy of doubtful value.

#304

Request Effect of administration of chloramphenicol to pregnant

female during first trimester.

Formulation CHLORAMPHENICOL and AMNION
or
PREGNANCY
COMPLICATIONS

Unwanted articles 1. Antibiotics in the tissue culture of human amnion (not adverse effects of chloramphenicol). 2. Chloramphenicol used therapeutically in a case of attempted abortion.

Solution CHLORAMPHENICOL/ADVERSE EFFECTS

#452

Request Complications (i.e., sequelae) of pancreatitis

Formulation PANCREATITIS and POSTOPERATIVE COMPLICATIONS
or
CHOLELITHIASIS

Unwanted articles Cases of postoperative pancreatitis (not postoperative complications following pancreatitis). Cases of pancreatitis and cholelithiasis in the same patient with no evidence of a relationship between the two. Cases of gallstones causing pancreatitis (i.e., the reverse relationship to the request).

Solution PANCREATITIS/SEQUELAE and POSTOPERATIVE COMPLICATIONS
or
CHOLELITHIASIS/ETIOLOGY

#457

Request Neurological complications of kidney diseases.

Formulation AMINOACIDURIA, RENAL
or
HYPERTENSION, RENAL
or
NEPHROSIS
or
NEPHRITIS
and
HEADACHE
or
COMA
or
DEPRESSION
or
NEUROLOGIC
MANIFESTATIONS

Unwanted articles 1. The neurological manifestations result from drug therapy or diagnostic procedure and not directly from the kidney disease. 2. Renal aminoaciduria in patients with cerebral disease, not kidney failure leading to cerebral disease.

Solution Kidney disease term/SEQUELAE

#462

Request Cerebral amyloidosis

Formulation AMYLOIDOSIS and CEREBRAL HEMORRHAGE

Unwanted articles Not amyloidosis of the brain, but an autopsy report on a case of macroglobulinemia with amyloid degeneration. The patient died of cerebral hemorrhage.

Solution There is no simple way of avoiding this failure by means of subheadings. However, as noted under #260, the coordination of two disease terms to express site of disease is a strategy of doubtful validity.

#473

Request Infections precipitating myasthenia gravis

Formulation MYASTHENIA GRAVIS and PITYRIASIS
or
PNEUMONIA

Unwanted articles The two conditions co-exist but there is no suggestion that the infection leads to myasthenia gravis.

Solution PNEUMONIA/SEQUELAE and MYASTHENIA GRAVIS/ETIOLOGY

#474

Request Intestinal atony or ileus following vagotomy or pyloroplasty.

Formulation PYLORIC STENOSIS and SURGERY, OPERATIVE
or
POSTOPERATIVE COMPLICATIONS
and INTESTINAL OBSTRUCTION

OR

VAGOTOMY and GASTROINTESTINAL MOTILITY

Unwanted articles 1. Cholelithiasis causing pyloric stenosis. 2. Pyloric stenosis and intestinal obstruction in patients following ingestion of chemical substances. (In neither of the above cases does surgery lead to intestinal obstruction). 3. Use of segments of the jejunum, together with vagotomy, in treatment of disabling post-gastrectomy symptoms; vagotomy does not lead to ileus or atony.

Solution 1. PYLORIC STENOSIS/SURGERY and SURGERY, OPERATIVE/ADVERSE EFFECTS and INTESTINAL OBSTRUCTION/ETIOLOGY

or

PYLORIC STENOSIS/POSTOPERATIVE COMPLICATIONS and INTESTINAL OBSTRUCTION/ETIOLOGY

(see discussion #474 under "false coordinations").

2. VAGOTOMY/ADVERSE EFFECTS and GASTROINTESTINAL MOTILITY

#481

Request Statistical correlation between the American Society of Anesthesiologists' physical status category and the intra- or postoperative death rate.

Formulation ANESTHESIA and MORTALITY

Unwanted articles In an evaluation of treatment of 2000 cases of massively bleeding gastroduodenal ulcers, certain general factors (e.g., medications, vitamins, diet, anesthesia, surgical procedure) influencing incidence, type and outcome of patients are discussed. There is no direct discussion of anesthetic death.

Solution ANESTHESIA/MORTALITY

#484

Request Complications following the use of oral contraceptives.

Formulation CONTRACEPTIVES, ORAL and STERILITY, FEMALE
or
MENSTRUATION DISORDERS

Unwanted articles Therapeutic use of estrogens in treatment of sterility and menstruation disorders.

Solution CONTRACEPTIVES, ORAL/ADVERSE EFFECTS, whereas the unwanted items would be indexed CONTRACEPTIVES, ORAL/THERAPEUTIC USE.

#513

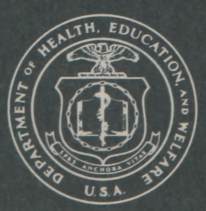
Request Effect of irradiation on the infectivity and structure of viruses.

Formulation RADIATION and PLANT VIRUSES

Unwanted articles Not the effect of radiation on viruses, but x-ray scattering used as an investigative tool.

Solution PLANT VIRUSES/RADIATION EFFECTS

U.S. GOVERNMENT PRINTING OFFICE : 1968 O—307-006



U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service