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 $\pm$ 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%

307-006 O-68—13

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.

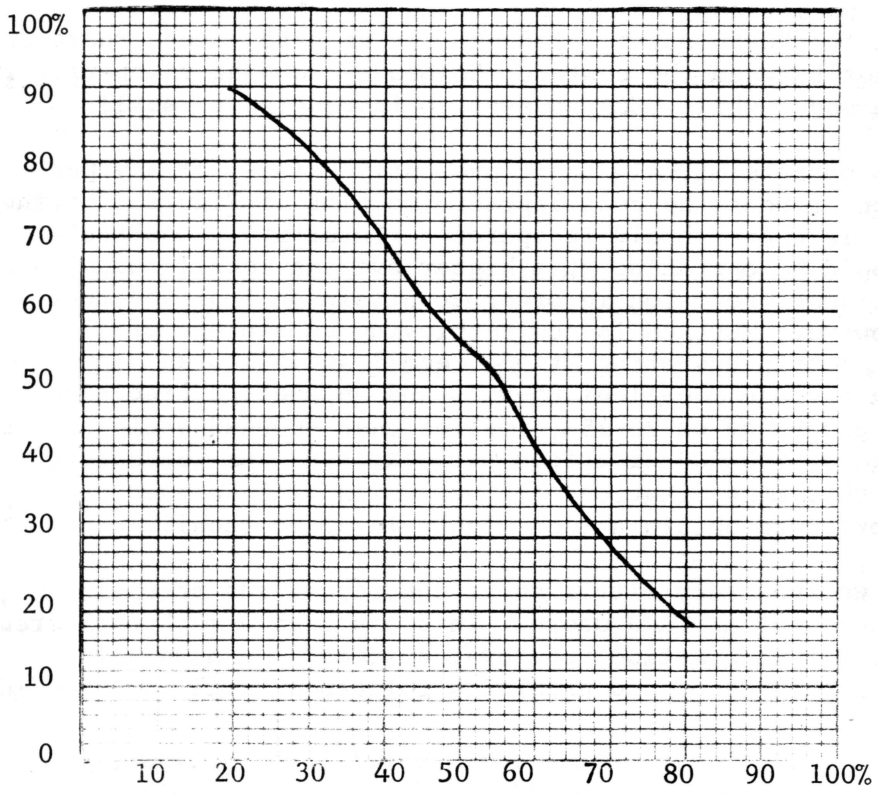
The fact that, on the average, MEDLARS is operating at 58% recall and 50% precision, indicates that, consciously or unconsciously, the MEDLARS searchers choose to operate in this general area. It would be possible for MEDLARS to operate at a different performance point on the recall/precision curve of 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 180 citations, of

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. W. L.

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.

307-006 O-68—14

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.