An Introduction to Medline Searching

Don R. Swanson
The University of Chicago

Preamble
The literature problem in biology and medicine

Part I
The algebra of sets and Venn Diagrams
-- a brief introduction
written for non-mathematicians. It is
limited to what online searchers should know.

Part II
Rules of the Game
The main types of search commands
available in PubMed, Ovid, and Dialog

Part III
Search Strategy
A systematic way of thinking about how to
search large bibliographic databases such
as Medline; controlling recall and precision.

Part IV
PubMed Puzzles
Five PubMed searches
that have surprising results

Acknowledgment:
This work has been supported by a collaborative grant to the
University of Illinois (PI: Neil Smalheiser, M.D., Ph.D.)
and The University of Chicago: 1 R01 LM07292-01,
Arrowsmith data mining techniques in neuro-informatics,
co-sponsored by NLM and NIMH, 6/15/01- 5/31/06.

The Literature Problem in Biology and Medicine

The immense size of the biomedical literature presents profoundly difficult search problems. The three largest electronic bibliographic databases that specialize in biology or medicine -- Medline, Embase, and BIOSIS (Biological Abstracts) as of 1996 jointly contained about 16 million unique records, each representing a published journal article or letter or report. Growth rate was about 1 million records per year derived from more than 12000 journals. At least 100 other electronic bibliographic databases also contain large amounts of life sciences material, including especially Scisearch (Science Citation Index) with over 13 million records in science, and several large agricultural databases. Even specialized literatures are enormous; cancer, for example, is today represented by 1.5 million records in Medline alone. Specialty articles are more widely dispersed among journals than is generally realized; a sample of 2000 articles on obstructive lung diseases, selected at random from 70,000 in identified in 1996 by a Medline search (today there are over 100,000) was found to be scattered among 650 different journals, with no one journal covering any substantial portion of this literature. We are entitled to suspect that keeping up with any single field is beyond human capacity, and that most of what can be found in a database search will not have been seen by the searcher when it was first published. More important, integration across fields is almost nonexistent and the fragmentation of knowledge has become the central information problem of science and scholarship. Online searching cuts across artifactual boundaries and so offers hope for the development of integrative mechanisms.

Online Access

Access to databases is provided by numerous vendors, each of whom offers choices among many databases and a highly useful command language for searching. Here we are concerned primarily with the National Library of Medicine (NLM) and the Medline database.

The internet can be an impediment to understanding online searching. Historically, online searching meant searching organized databases using a comnputer. Today it tends to be confused with searching the internet, to which it bears only a slight resemblance. The internet does provide a communication channel that permits the user to link up with the online search services, including NLM, but it is not a mechanism for searching Medline.

Online searching is the art of forming and combining sets, and so we begin with a brief introduction to the algebra of sets.

Part I

The Algebra of Sets and Boolean Search Statements: a brief introduction

1.1 Sets can be formed of almost anything, but there are some constraints, -- expressed by the following rules:

1. Identify at the outset a "universe" of things under consideration -- often called the "elements" of the set. To get right to the point of Medline searching, our universal set will be the set of all records in Medline, each record being represented by a unique 8-digit number, called a Unique Identifier (UI). Although our sets can be thought of as sets of biomedical articles, or their corresponding records in Medline, it is the UI number that is used internally by the computer in combining sets, and is free of ambiguity.
2. For the universal set, as well as any other set, there must be no ambiguity about whether something is or is not a member of the set. Set membership is not a matter of opinion; it is a binary variable -- if the issue is whether a particular UI is a member of some set, the answer must be "yes" or "no".
3. Each element is distinct and distinguishable from each other element; thus no element can be a member of a set two or more times. If some element, such as a UI number, is listed twice, duplicate listings are ignored or eliminated. (The only way the same article can sneak into Medline twice is to get two different UI numbers.)
4. The order in which elements are listed doesn't matter.

1.2 Examples of sets

In these notes, all examples will be constructed using 3-digit elements as set members -- to save the trouble of typing 8-digit UIs.

Set notation: a common way to designate a set is a pair of brackets, within which each element of the set is listed:

{962 224 777} contains 3 elements: 224, 777, and 962.

{962 224 777 224 962} would be the same set; normally such repetitions cannot occur, but if they did, they would simply be eliminated; the set would not be abolished, only the duplicates would be abolished.

{777 962 224} is also the same set; sequence doesn't matter.

{} has no members and is called a null set
{ --- all UIs in Medline --- } is the universal set

1.3 Boolean operators for online searching

An alternate notation that reflects the relationship to a Boolean algebra is to use uppercase letters to designate sets;

A = {962 224 777}
I = { --- the universal set --- }
O = {}, the null set

A useful extension of this notation applies to online searching: We can think of A as a search word or search term that creates a set of record numbers. Thus A might stand for the set of all records that contain the word "apples".

The complement of set A, called A', consists of all members of the universal set that are not members of set A. If set A is defined as above, then A' = { -- all UI numbers except 962 224 777 -- }
Notice that the complement of the universal set is the null set and vice versa: I'=O and O'=I

The number of members of set A is denoted n(A).

1.4 Relationships between sets; five crucial definitions:

1. Two sets are equal if they have the same elements.
2. The intersection of two sets is the set of all elements common to both:
   intersection of {224 962 777} with {225 963 777} = {777}

Intersection in Boolean notation is denoted by an operator "AND" so that the intersection of set A with set B is called

A AND B

Thus if A={224 962 777} and B={225 963 777} and C={777}
then A AND B = C
Notice that: A AND B = B AND A
A AND I = A I AND A = A
A AND O = O O AND A = O
From the definition of complement, it follows that A AND A'= O

3. The union of two sets is the set of all elements in either set:
   the union of {224 962 777} with {225 963 777} = {224 225 962 963 777}
   Union in Boolean notation is denoted by an operator "OR"

The union of set A with set B is: A OR B
From the definition of complement, it follows that A OR A'=I

4. Negation makes use of the complement:

A AND B' is equivalent to A NOT B

5. A is a subset of B if all the members of A are also members of B, and it follows that the number of members of set A is less than the number of members of B.

This will be written here as A < B and it follows that n(A) < n(B)
This subset relationship also implies that IF A < B, then:

A OR B = B A AND B = A

6. To "combine" sets now can be given a special meaning, namely to use the relationships 2, 3, 4 defined above.
The earlier statement: "Online searching is the art of forming and combining sets", is now more explicit.

1.5 Ordering and grouping of search statements

With strings of AND, or strings of OR, the order of sets doesn't matter, nor does it matter how they are grouped (associated):

A AND B AND C is the same as C AND (A AND B) etc
A OR B OR C is the same as (C OR A) OR B etc

But in mixed statements, the order does make a difference:

A OR B AND C will be interpreted differently by different implementations of search command languages -- PubMed processes from left to right and Dialog assigns an order of precedence to the operators, with AND being processed before OR. Rather than trying to remember who does what, such statements should always be grouped using parentheses to reflect the users intent. (A OR B) AND C is not the same as A OR (B AND C), but neither is ambiguous -- operations within parentheses will always be given precedence. The difference between these two relationships may be easily envisioned using a Venn diagram (Fig. 1).

1.6 The Venn Diagram as a "Container" Model

Venn diagrams provide not only a way to visualize set relationships, but a sound way to conceptualize sets and to solve problems in the

forming and combining of sets.

Draw a large rectangle to "contain" the universal set. Then, within the rectangle, draw a circle to represent set A. The interior of the circle "holds" all the elements of A; exterior to A, but within the rectangle, are all members of the complement, A'. The intersection of two circles (A, B) represents or "holds" the elements of the intersection of the two sets. The union of the two sets consists of all elements inside of either or both circles.

Circles may be replaced with any closed figures; indeed, when drawing more than three figures, it may not be possible to represent all intersections without distorting the circles.

1.7 Problem solving with Venn diagrams: examples

Show that the number of records in the union A OR B is calculated by adding the number in A to the number in B and subtracting the number within the intersection A AND B. Fig 2.

The results of an online search for the following words and combinations of words are shown by the numbers of the left column, which represent the number of records found:

150 spreading
1087 cortical
1549 depression
17 spreading AND cortical
35 spreading AND depression
25 cortical AND depression
17 spreading AND cortical AND depression

Draw a Venn Diagram using 3 circles that intersect, and label each of the seven enclosed spaces with the appropriate number. How many records have the words "spreading" and "cortical" but not the word "depression"? Fig. 3

Part II The Rules of the Game - in brief -

2.1 Medline design is MeSH based

There are a number of implementations of Medline, from various suppliers of online search services, and the command language and search capabilities (i.e. the rules of the game) are not all identical. Those of interest here all provide access to the Medline database, which consists of just over 10 million records corresponding to published biomedical articles, reviews, letters, and reports. Record formats differ somewhat and all suppliers provide a variety of formats, but the most important feature they have in common -- other than the database itself -- is the indexing supplied by the National Library of Medicine.

Medline is designed basically to provide access by means of human-assigned medical subject headings (MeSH), and so some further comments about this are in order; it is, after all, the heart of the system, even though not classified anatomically under A7.541.

2.2 MeSH and the process of indexing

The MeSH system began to evolve, along with the printed Index Medicus at the turn of the century, and has been steadily growing. It is under continual review and modification in response to how medical and lexicographical experts perceive the changing needs of biomedical practitioners and researchers.

The process of indexing each biomedical article to be included in the database has also been evolving. Although this process has made increasing use of computerized aids, it is probably fair to say that it has always entailed a careful reading of each article in its entirety by well-trained indexers who follow prescribed rules and whose work is supervised and reviewed. The indexers who apply the MeSH terms and subheadings to each article are of course not infallible; inter-indexer consistency has been studied and found wanting. But the nature of the errors should also be understood; most are errors of omission rather than commission, and of subheadings more than main headings. Indexers are instructed to index a topic only if it is substantively treated. The indexing of a minor mention of a topic is not warranted. But the question of substance or importance being a matter of judgment, differences of opinion are bound to occur. On balance, the index terms lead to much better access than the assumptions a searcher makes about what words the author might use in the free text of the title and abstract. Natural language is sufficiently rich to defeat most attempts to anticipate exactly how a particular idea or concept will be expressed in words.

The purpose of the MeSH structure is to provide a standardized language, hierarchically-organized, that will bridge the gap between the natural language of the user and the author. Even though the MeSH structure and terms are searchable online, the nature of the system can best be appreciated by examining the printed product published (and updated annually) by NLM, and by using it in

the construction of search statements.

2.3 The printed MeSH reference manuals

The printed MeSH manual is divided into three parts, listed here in the most logical order of use by a searcher, each part providing entry information for the next part.

1. Permuted Medical Subject Headings
   This volume is organized alphabetically by each word that occurs in any medical subject heading; under that word, it lists all the subject headings in which the word occurs. It also includes words that do not occur in any subject heading, but which serve to refer the user to the correct subject heading entry. The permuted index serves as what is often called an "entry vocabulary".
2. Medical Subject Headings -- Annotated Alphabetic List
   This volume lists all of the official headings, in alphabetic sequence and with brief comments on usage, and the hierarchical codes associated with that term that provide entry into the tree structure. This volume also includes the list of subheadings (of which there are about 80) that can be attached to main headings. There are many cross-references.
3. Medical Subject Headings -- Tree Structures
   This volume is not alphabetic, but is organized by subject using a decimal code that can reflect up to about ten levels. Any one mesh term may appear in multiple parts of the classification tree. Whenever an "exploded" search is conducted, this tree structure can be used to show exactly what MeSH terms the search will cover.

For any serious searching, the above three volumes are almost indispensable, attempts by the providers of online search services to make the whole process "transparent" to the user notwithstanding.

2.4 The MeSH hierarchy and the meaning of "explosion"

To "explode" a MeSH term is to expand a search to cover not only the starting term, but everything subordinate to it in the MeSH tree. Indexers are instructed to use only the most specific terms applicable to the article being indexed; they do not insert into the index any term higher in the tree unless the article itself treats the material at that higher level. This process gives the user some choice in the

level at which to conduct the search -- i.e. to explode or not to explode.

2.5 Categorization of search commands

The three most important categories of search commands are to search, to combine, and to display -- that is, to form sets, to combine sets, and to display sets or records. A set refers to a set of Medline records.

The earliest systematic online search command language was probably that designed by Dialog around or about 1970. The above types of command were explicitly given the names: select (later changed to find), combine, and type or display.

For the most part, present systems do away with command names altogether, which may look simpler to the user but which obscures important conceptual distinctions. In most cases, just entering a word, phrase, or term in the search window implicitly invokes a search command. Entering AND, OR, or NOT implicitly invokes a "combine" command, obscuring the fact that this command does not search the database, but operates only on already formed sets whose elements are represented by UI numbers. Display commands operate on already formed sets, and can display specified elements and specified parts of the record. There are also commands that can apply to a series of searches and are invoked by setting "limits".

2.6 Search qualifiers

Every supplier of online services provides some means for the user to choose which field or fields to search, and which subheadings to attach to each chosen MeSH term.

2.7 A few differences between Ovid and PubMed:

1. PubMed uses brackets to enclose field qualifiers, Ovid uses periods;
2. MeSH terms are MH in PubMed and sh in Ovid.
3. Set numbers in PubMed have a preceding #.
4. Boolean operators in PubMed must be uppercase.

PubMed		Ovid
Thus neuroglia[MH]	corresponds to	neuroglia.sh.
#1 AND #2	corresponds to	1 and 2

5. TW in PubMed and tw in Ovid both stand for text-word
   but correspond to somewhat different fields; .tw. in ovid means
   titles and abstracts, and corresponds to PubMed's [Title/Abstract].
   TW in PubMed includes the main vocabulary of MH terms and what is
   called the entry vocabulary, in addition to the title and abstract.
6. For a phrase search, PubMed requires quotes around the phrase

and will match only what it finds in its index.
Ovid will search all phrases in the title and/or abstract.
7 Most important, PubMed explodes all medical headings by default use [MH:noexp] to shut off, while Ovid requires the command: exp

Before an informed choice of a search field can be made, the user should understand the nature of the records being searched. A Medline record has numerous fields; a format that labels each field with a two-letter code at the left margin provides the clearest illustration of what a record is like:

2.8 Example of a PubMed Record

UI - 89116243
PMID- 2536517
DA - 19890301
DCOM- 19890301
LR - 20001218
IS - 0002-9343
VI - 86
IP - 2
DP - 1989 Feb
TI - Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study.
PG - 158-64
AB - PURPOSE: The ingestion of omega-3 fatty acids could benefit patients with Raynaud's phenomenon because, among other effects, these fatty acids induce a favorable vascular response to ischemia. The aim of our study was to investigate, in a double-blind, placebo-controlled manner, the effects of fish-oil fatty-acid dietary therapy in patients with rheumatic disease. PATIENTS AND METHODS: Thirty-two patients with primary or secondary Raynaud's phenomenon were randomly assigned to olive-oil placebo or fish-oil groups. Patients ingested 12 fish-oil capsules daily containing a total of 3.96 g eicosapentaenoic acid and 2.64 g docosahexaenoic acid or 12 olive-oil capsules and were evaluated at baseline and after six, 12, and 17 weeks. All patients ingested olive oil between Weeks 12 to 17. Digital systolic blood pressures and blood flow were measured at room air and water baths of 40 degrees C, 25 degrees C, 15 degrees C, and 10 degrees C using strain gauge plethysmography. Onset of Raynaud's phenomenon was timed with a stop watch and defined as plethysmographic evidence of cessation of blood flow and blood pressure in the study finger. RESULTS: In the fish-oil group, the median time interval before the onset of Raynaud's phenomenon increased from 31.3 +/- 1.3 minutes

baseline to 46.5 +/- 2.1 minutes at six weeks (p = 0.04). Patients with primary Raynaud's phenomenon ingesting fish oil had the greatest increase in the time interval before the onset of the condition. Five of 11 patients (45.5 percent) with primary Raynaud's phenomenon ingesting fish oil in whom the phenomenon was induced at baseline could not be induced to develop Raynaud's at the six- or 12-week visit compared with one of nine patients (11 percent) with primary Raynaud's ingesting olive oil (p = 0.05). The mean digital systolic pressures were higher in the patients with primary Raynaud's phenomenon ingesting fish oil than in patients with primary Raynaud's ingesting olive oil in the 10 degrees C water bath (+32 mm Hg, p = 0.02). CONCLUSION: We conclude that the ingestion of fish oil improves tolerance to cold exposure and delays the onset of vasospasm in patients with primary, but not secondary, Raynaud's phenomenon. These improvements are associated with significantly increased digital systolic blood pressures in cold temperatures.

AD - Division of Rheumatology, Albany Medical College, New York 12208.
FAU - DiGiacomo, R A
AU - DiGiacomo RA
FAU - Kremer, J M
AU - Kremer JM
FAU - Shah, D M
AU - Shah DM

LA - eng
PT - Clinical Trial
PT - Journal Article
PT - Randomized Controlled Trial
CY - UNITED STATES
TA - Am J Med
JID - 0267200
RN - 0 (Placebos)
RN - 1553-41-9 (5,8,11,14,17-Eicosapentaenoic Acid)
RN - 25167-62-8 (Docosahexaenoic Acids)
SB - AIM
SB - IM
MH - 5,8,11,14,17-Eicosapentaenoic Acid/*therapeutic use
MH - Adult
MH - Aged
MH - Blood Pressure
MH - Clinical Trials
MH - Docosahexaenoic Acids/*therapeutic use
MH - Double-Blind Method
MH - Female
MH - Fingers/blood supply
MH - Human

MH - Male
MH - Middle Age
MH - Placebos
MH - Prospective Studies
MH - Raynaud's Disease/*diet therapy/etiology/physiopathology
MH - Regional Blood Flow
MH - Scleroderma, Systemic/complications
EDAT- 1989/02/01
MHDA- 1989/02/01 00:01
PST - ppublish
SO - Am J Med 1989 Feb;86(2):158-64.

2.9 The PubMed help file as a user's guide

An explanation of each field can be found in the PubMed "help" file, as well as a far more detailed description of the search tools available. A rather good user's guide can be created by downloading and printing this file. Just reading it as a book is extremely helpful for any searcher who must first figure out what to ask before searching for the answer. It helps just to flip through it to see how it is organized and what it covers.

2.10 Phrases and Proximity searching

There is one important class of commands PubMed does not provide.

Ovid, Dialog, and other suppliers (not including PubMed) keep information on the location or address of each text word in the electronic record, and so are able to provide the searcher with the option of specifying a phrase or string of words that appear next to or near each other within the title or abstract of a record. In Ovid, for example, a search for information adj2 retrieval.ti. will find 278 records with the specific phrase "information retrieval" in the title, and 96 more records in which one or two words may intervene between the two given search words, and the search words may be in either order -- one record having the title "Retrieval of relational information: a role for the left inferior prefrontal cortex." Dialog gives the searcher an additional option of specifying the order of occurrence -- which is often valuable.

PubMed attempts to compensate for the lack of locational or proximity information by allowing a phrase search for any phrase that it can match to an entry in a stored list of phrases. (To force a phrase search, put the phrase in quotes.) But it does not reject (non-quoted) phrases that do not match -- instead it tries to provide an approximation or alternative to the user's requested search. The searcher can find out how the search was actually conducted by clicking on a button called "Details". The results are sometimes surprising. More will be said of this in the discussion of search strategy in Parts 3 and 4.

Part III Search Strategy

Strategies and tactics for searching Medline
(and other large bibliographic databases).

3.1 Advice from Japan

Long ago I bought a bicycle, shipped unassembled from Japan. An enclosed list of 75 instructions began:

1. Before assembling Japanese bicycle, compose yourself.

That proved by far the best piece of advice on the list. It's also a good start for thinking about information retrieval (IR) -- the initials not to be confused with Immunoreactive, Infrared, Insulin Resistance, Insulin Receptors, or International Relations, all of which can be found in an online search for IR.

3.2 IR Theory?

There is no good theory of IR. That's a bit strong a statement, and some would hotly contest it. But, as Bill Cooper, a logician who became a leading IR theorist once said "deep down, it's shallow." So that's why we talk about strategies here rather than theories.

Strategies for searching can usefully be organized around two categories -- searching of the natural language text of titles and abstracts, and searching of standardized languages, also called controlled vocabularies, such as MeSH. Neither is very good, but they can complement one another to produce results that are better than either one alone. It is possible to partly automate the process of translating from a natural-language search request to the appropriate MeSH terms, but not to the extent that the searcher can afford to count on it. In short, understanding MeSH is important to searching Medline.

3.3 The "all but only" goal; recall and precision

The goal of an online search is to find all that is useful, in the context of your search, but only what's useful -- abbreviated as the "all but only" goal. No system or strategy has ever been designed that can consistently do both. For any given search, if a series of strategies is attempted that vary in how much they retrieve, in general, the closer you get to "all", the more junk you have to wade through, and the harder you try to filter out all the junk in a quest for "only", the more you lose of what you want. Designing a good strategy is the art of finding the best compromise.

There's jargon in all fields, and IR researchers refer to the above phenomenon as the "recall-precision tradeoff", where, for any given search:

recall = percent retrieved of the putative total of useful stuff (the latter being

generally unknowable), and
precision = percent relevant (or useful) of everything you retrieve.

Evaluating results is a labor-intensive process that requires particiation of the person who asked the question, for it is only the questioner who can decide ultimately what qualifies as a useful answer.

3.4 Tests conducted long ago that still have currency

Extensive realistic tests of Medline searching -- involving about 300 test questions -- were conducted by Wilf Lancaster in 1968, leading to an overall average of close to 50% recall (estimated by retrieval of previously-known relevant documents), and 50% precision, but with individual search results that varied all over the map from 0% to 100%. Any searcher, with or without being aware of it, inevitably must select a search strategy that is geared to some particular compromise between recall and precision. There is a large difference between good compromises and bad compromises, but high variability from one search to another is dependable.

3.5 The independent variable: breadth

An important variable on which recall and precision depend is that of the "breadth" of a search. I will take "breadth" to mean essentially the total number of records retrieved. (There is another sense in which "broader" is taken to mean moving up in the hierarchy of MeSH terms; I shall make it explicit when that meaning is intended). Every bibliographic online search system recognizes the crucial importance of letting the searcher know how many records were found. This single number, instantly displayed by the system immediately after every search, often tells you what to do next to improve the search. Think about the difference between learning that you found 1.4 million records and learning that you found 3 records.

3.6 The Recall/Precision tradeoff

As a search is made broader, more is retrieved and therefore more relevant or useful records are retrieved -- along with more that are not useful. Thus recall always increases when a search is broadened. Precision may increase or decrease -- but if an average over many questions by many searchers is taken, what emerges is a tendency for precision to decrease as breadth is increased. There are number of specific search strategies that increase recall, and a similar number that increase precision.

3.7 ***R Strategies that increase recall (R)

R1: Liberal use of the Boolean operator OR, as in
A OR B OR C OR D OR E ----------
String enough ORs together and you will get everything.

R2: The "explode" command (applied to specific MeSH term, MH) will generate all records indexed by MH, or by any term subordinate to it in the MeSH tree. (Thus it implicitly has many ORs built into it.) PubMed automatically explodes all MeSH-term searches, unless you tell it not to. Ovid requires an explicit command, as in: exp virus diseases.

R3: Moving upward in the (exploded) MeSH hierarchy will broaden a search (not necessarily true of unexploded terms). For example, Parkinson Disease, with 23,000 Medline records, belongs within the broader category of Neurodegenerative Diseases, but the latter has only 2000 Medline records because indexers are instructed to assign only the most specific terms, unless the article also discusses the broader level. Exploding neurodegenerative diseases yields 114,000 records.

R4: The more Fields you choose to search, the more records you will retrieve. PubMed will default to All Fields, unless the term searched is specifically attached to a field designator (such as the MH in Parkinson Disease[MH]) or the searcher clicks on "Limits", then on "All Fields" to choose a specific field for all search terms. Ovid requires the searcher to specify the field for each search term. If the field is not specified, it displays options for "mapping" to the what it thinks is the correct subject heading and subheading.

R5: Truncate search terms (using "*" in PubMed or "$" in Ovid) This feature automatically generates a string of ORs, that connect all search terms that begin with the specified string. Use with caution. Be sure you know exactly what you are asking for. On PubMed, click Preview/Index, enter the term in the window and click Index. Best to avoid if in doubt. Generally, don’t even think about truncating very short common words, like diet*. (Truncating xxxy* is ok.)

3.8 ***P Strategies that tend to increase precision (P)

P1: Liberal use of the Boolean operator AND, as in
A AND B AND C AND D AND E ----------
String enough ANDs together and you will get nothing.
A corollary to this idea is to search for long
phrases.

P2: In searching MeSH terms, shut off the automatic explosion
in PubMed, as in: neurodegenerative diseases[MH:noexp]

P3: Moving downward in (exploded) hierarchies will decrease
the number of records retrieved, the converse of
R3 above.

P4: Limiting a search to title words is often a
powerful way to increase precision -- it leads
to fewer records that will also tend to be more
relevant and central to the concepts searched.

P5: In searching MeSH terms, add subheadings: brain edema/etiology

P6: Free-text phrase and proximity searching in Ovid.
The two word phrase "A B" will be narrower, i.e. produce
fewer records than "A AND B". Intermediate between "A B"
and "A AND B", is the (Ovid) proximity search "A adjn B"
where "n" is the maximum number of intervening words.

P7: Rare or unusual words and phrases, if appropriate to
a search, can be effective in producing a focused,
high-precision, output. (If the word is necessary as
well as appropriate, then it may also produce high recall.)

3.9 Strategic techniques, ideas, or moves

Next is presented a number of strategic "ideas" -- SI-1, SI-2, SI-3, etc. which means that these are neither "rules" that you must follow (as in the "rules of the game") nor are they as general as "principles" -- they are simply useful moves to think about.

SI-1 What may be the best way to start any new search is to
begin with a high-precision title search in order
to determine whether at least one or more highly

responsive articles exist in the Medline database.
If an article is found that is close to what you're looking for, then study how it is indexed! This is often the best way to find the right MeSH terms as a basis for conducting a new search that may lead to far higher recall without excessively sacrificing precision. Or, MeSH terms can be used to further improve precision by forming an intersection with a title-word search.

SI-2: Always examine at least the titles of a dozen or so records that are retrieved in any search you conduct; Look for the context in which search terms are used. This suggestion will be of greatest value if you have based the search on title words. If you have not done so, then examine also abstracts and MeSH terms. Be sceptical always. You will often encounter surprises in how words are used in contexts that may not have occurred to you. Then, of course, revise your search strategy as needed.

SI-3 When examining the results of a multi-term search it is more informative to break the search into single steps, so if you do encounter any surprises, the cause will usually be instantly clear. A common surprise is to get "0" because of a typo in your entry, or perhaps you see a posting number that is 10 or 100 times greater than you might reasonably have expected. This suggestion is especially important in PubMed, where the "details" of what it did may be almost indecipherable for a complex search. Leave all combinations to the end, where they can be more lucidly constructed by referring just to set numbers.

That is, if (A OR B) AND (C OR D) AND E yields 0 or any other unexpected result, the reasons can be more easily deduced from the following version of the same search:

set#	term	#records found	comments
1	A	421	
2	B	85326	(is this reasonable?)
3	C	101	
4	D	66	
5	E	1	(is this reasonable?)

6 1 OR 2 85326 (showing, perhaps unexpectedly, that A is a subset of B)
7 3 OR 4 122
8 6 AND 7 0 (is this reasonable?)
9 5 AND 8 0 = (1 OR 2) AND (3 OR 4) AND 5

SI-4 When you have followed the foregoing suggestions, next create two practice versions of your search, one that attempts high precision and one that goes for high recall. This exercise often can guide you to a reasonable compromise.

A bit more jargon that is uncommon, but helpful here:
Definition:
"searchonym" = synonyms, inflectional variants such as singular and plural, etc., and other related terms that tend to serve equivalent functions in a search

SI-4 Try to formulate your search in terms of a sort of generic strategy that can be called a "canonical" form, as follows:
(A OR B OR other searchonyms ---) AND
(C OR D OR other searchonyms ---) AND ----- etc.
In general, words, phrases, or MeSH terms should be connected with "OR" only if they have similar enough meanings to serve essentially the same purpose in a search. The operator AND is then used to connect entirely different groups of searchonyms that should have one or more members present in any relevant document.

3.10 When to stop

In sum, online searching of bibliographic databases is a highly imprecise interactive trial-and-error process in which you make your best initial guess, do the search (usually starting with a title-word search), then assess the number of records that were found, have a look at a dozen of them -- especially looking at how they were indexed, and then revise the search by incorporating index terms, and try again. Repeat this process of forming and combining sets until satisfied or fed up, whichever comes first.

Part IV

PubMed Puzzles

Each of these puzzles is based on the assumption that you do what PubMed virtually invites, and is designed to respond to, -- enter a phrase without field qualifiers or quotes, -- i.e. in free normal natural English language. (Exception: if you do use Boolean operators, (AND, OR), use uppercase). In general, however, PubMed attempts to take over the task of translating the natural language of the question to a proper MeSH search statement.

In each puzzle example, the entry phrase or combination is supplied, along with the number of records that are found.

4.1 Puzzle #1: biological agents OR biological weapons --- 43000

How do you decide whether this number is reasonable?

First step is to look at the records.

Viewing the first dozen or so showed that none seemed to be relevant; at least the search phrase could not be found anywhere in the record.

What could be wrong?

Click on "details":

A revelation: the search actually executed was:

(biological AND agents) OR (biological weapons)

Diagnosis: PubMed was unable to find the phrase "biological agents" in its vocabulary so, instead of admitting this awkward fact, it gave you what IT decided was the next best thing without finding out if that was a good approximation to what you wanted. (It helps to some degree here to place each phrase of the search statement within quotes; this elicits the warning that PubMed could not identify the phrase, but then it still gives the same results as above.)

One way for the searcher to find out how good it was is to try the search in Ovid, which permits free-phrase searching (no vocabulary match required, because location information for each word in each record is kept).

Repeating the search as a text word in Ovid yielded:

1 biological agents.tw. 581
2 biological weapons.tw. 236
1 or 2 789

Viewing a sample of the record titles, all appear to be relevant.

Thus PubMed reported 55 times as many records as Ovid and none among those examined seemed to be relevant. Clearly the set of records with the phrase "biological agents" is a very small subset of all records with both words (581 out of 43000) or 1.4%, -- hardly a good approximation. ("biological weapons", -- entered without quotes, involved similar issues but relatively few (450) records; in

this case, forcing a phrase search by using quotes was successful, leading to 248 records).

The first lesson to be learned here is that it is not impossible to be led by PubMed into a dramatic search disaster if you fail to examine the "Details".

There is also a second, but not quite so obvious, lesson. The magnitude of the disaster was aggravated by the failure to put any constraints on the PubMed search field, allowing it by default to search "All Fields". Certainly it is plausible to limit a text search to titles and abstracts (as was done in Ovid), even if you accept the fact that PubMed cannot handle the phrase. Hence try:

biological[title/abstract] AND agents[title/abstract]
This yields, instead of 43000 records, only 8464
-- i.e. 15 times as much as what you asked for
instead of 55 times as much, but still mostly irrelevant.

Third, but obscure, lesson:

Do not use PubMed's [TW] ("text word", peculiarly defined) unless you know exactly what you are doing; it puts you back in the soup with 39000 records.

Bottom line remedy: switch to Ovid-Medline or Dialog-Medline for this kind of search. There seems to be no remedy within PubMed.

4.2 Puzzle #2 chemical agents OR chemical weapons -- 49512
whereas an Ovid text search yielded -- 2025
Ovid appears to be correct.
This is just a variant on puzzle #1
-- same issues raised; same diagnosis

4.3 Puzzle #3: postconception contraceptives -- 997
not obviously surprising, but next try
a slight variation that would seem to
mean exactly the same thing:
postconception contraceptive agents -- 34779

Diagnosis?: Perhaps PubMed simply could not make the leap from contraceptives to contraceptive agents. But, click on "Details" for the first search to see that it did cope with that leap:

The search actually executed was --
postconception[All Fields] AND
(contraceptive agents[MH] OR contraceptives[TW])

So now click on "Details" for the second search:
The search actually executed was -- Abortifacient agents
-- and the result was 35 times as many records

Comment: What PubMed thought you meant by the first search was not the same as what it thought you meant by the second search, which most human observers would have thought was identical to the

first search.
What does Ovid think of all this? It thinks there is no such thing:

postconception contraceptives.tw. -- 0
postconception contraceptive agents.tw. -- 0

-- at least, no such phrase occurs in any title or abstract in the Medline database. Moreover, Ovid's mapping function for the first phrase leads to Contraceptives but not to Abortifacient Agents. For the second phrase it leads nowhere.

On the assumption that the intent of the searcher is adequately met by the literature on Abortifacient Agents, PubMed gave a better first response than Ovid in this case -- among the 997 records to which the puzzle phrase (i.e. postconception contraceptives) led, 315 were also indexed with Abortifacient Agents; hence the searcher might have been led to them, but perhaps with the mistaken idea that there were only about 300 rather than 35000.

The lesson again is: Details, Details, Details.

4.4 Puzzle #4: induced dementia -- 3804
whereas an Ovid text search yielded -- 38
Examining the records can confirm that Ovid is correct, and PubMed results are 99% irrelevant.
Details of PubMed showed the actual search to be:
induced[All Fields] AND (dementia[MH] OR dementia[TW])

Again, the diagnosis is clear enough from the Details, -- the word "induced" occurred somewhere in the record, but in most cases not paired with "dementia". There is no apparent remedy within PubMed.

4.5 Puzzle #5: alcohol diseases OR alcoholic diseases -- 31490

"Details" again showed the replacement of phrase contiguity with AND, but no subject headings were found and so the main issue here is quite different. "If all else fails read the instructions" is a good maxim at this point -- in particular read the printed MeSH volume "Permuted Medical Subject Headings", which is in alphabetic order. Look up both alcohol and alcoholic to get a list of subject headings that contain that word, then pick out those that are diseases, to get:

Alcohol Amnestic Disorder
Alcohol-Induced Disorders
---- etc ---------

-- there are dozens more, but stop right here and check the list of

subordinate terms. This is done by going to the Annotated Alphabetic List, looking up "Alcohol-Induced Disorders" to get the classification code C21.739.100.87+ which in turn is used as an entry point to the Tree Structures volume, where one notices first that, one step higher, is the code for the slightly broader term

Alcohol-Related Disorders C21.739.100

-- subordinate to this code, one finds:
Alcohol-Induced Disorders
Alcohol-Induced Disorders, Nervous System
Alcohol Amnestic Disorder
Korsakoff Syndrome
Alcohol Withdrawal Delirium
Alcohol Withdrawal Seizures
Alcoholic Neuropathy
Cardiomyopathy, Alcoholic
Fetal Alcohol Syndrome
Liver Diseases, Alcoholic
Fatty Live, Alcoholic
Hepatitis, Alcoholic
Liver Cirrhosis, Alcoholic
Pancreatitis, Alcoholic
Psychoses, Alcoholic
Alcoholic Intoxication
Alcoholism
Wernicke Encephalopathy

An exploded search of Alcohol-Related Disorders would get all of the above, and yield, for both PubMed and Ovid, approx: -- 65000

The list may be of help in narrowing down the search, just in case, for example, the searcher really meant to ask only for alcoholic diseases of the nervous system.

An exploded search of Alcohol-Induced Disorders, Nervous System would yield, for both PubMed and Ovid, approx: --- 2500

Note that the 65000 undoubtedly is higher in precision (and more obviously in recall) than the original 31490, because the latter represents only the conjunction of diseases with alcohol(ic), and not necessarily alcohol(ic) diseases -- i.e. diseases related to or induced by alcohol. PubMed attempts to match the search statement to the correct MeSH terms, but in this case failed to find any of the above listed terms, all of which appear to be appropriate.

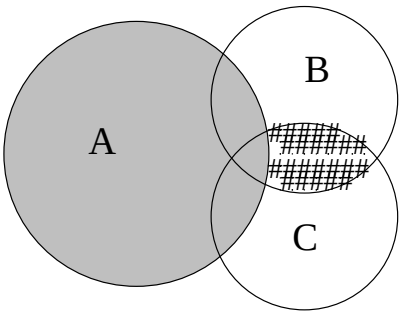
The lesson here is very general -- before conducting a search, find the best subject headings. Do not rely on PubMed's attempt to do so automatically. Having immediately available and learning to use the three MeSH printed

volumes is a prerequisite, and indispensable. But there are other valuable techniques, the most important of which is to find a few articles known to represent just what you are looking for, then examine the subject headings (MH tags) assigned to them.

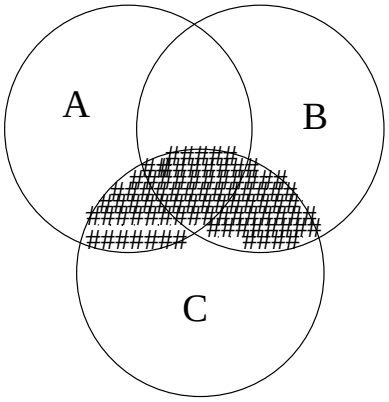
4.6 A concluding note: PubMed search tactics

It should be understood that the foregoing puzzles were selected for the lessons they provide; and not to suggest that all or even most PubMed searches are really that bad. One general implication however is that all unqualified or natural-languages searches of PubMed add a layer of complexity by requiring the searcher to figure out, if possible, what PubMed actually did. To bypass some of this layer of complexity or mystery a few tactical "rules of thumb" (RT) might be valuable, especially for inexperienced searchers:

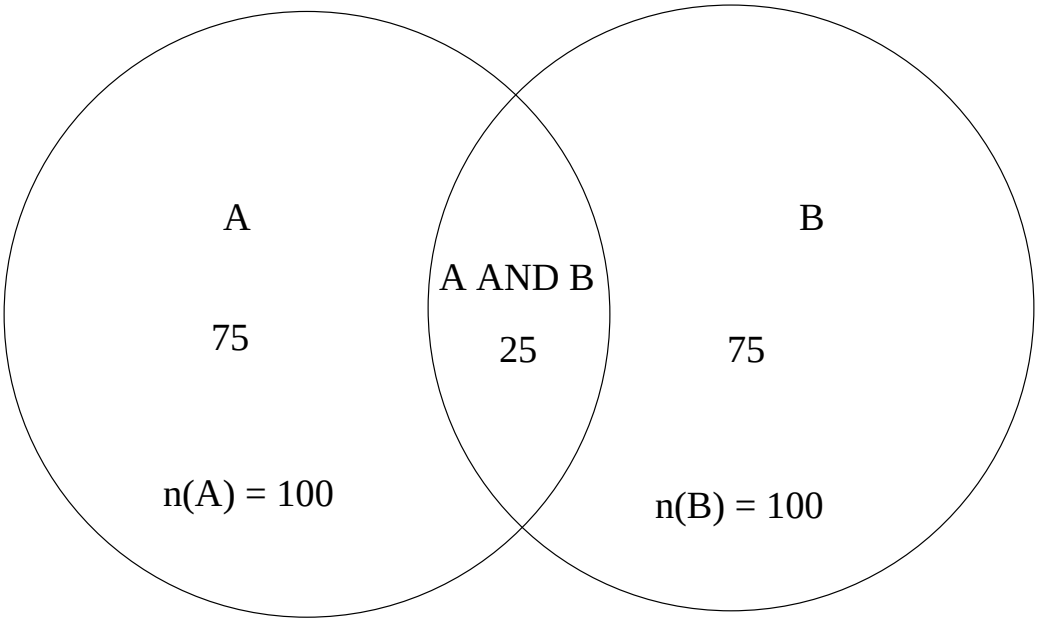
RT-1 Always view "Details".
RT-2 Always specify fields, using either bracket [] notation, or by setting a "limits" choice, -- i.e. avoid [All Fields].
RT-3 Almost always avoid [TW] -- at least until you understand what it does -- in general, substitute [Title/Abstract].
RT-4 Putting phrases in quotes is a good idea because it will evoke a warning from PubMed when it cannot find the phrase, so you know to look in the Details for how AND was applied.
RT-5 Finally, if Ovid is available to you, double-check your strategy by first developing it on Ovid, even if you then switch to PubMed. Compare the postings numbers (they will usually be slighly different) to flag any major mysteries. Be aware that PubMed explodes all MH terms unless told not to [MH:noexp], whereas Ovid requires the command: exp
RT-6 For any exploded MeSH term, examine the printed Tree Structures so you will know what you are getting from the explosion, namely everything subordinate to the term exploded. This information might help you narrow down the search.



A OR (B AND C)

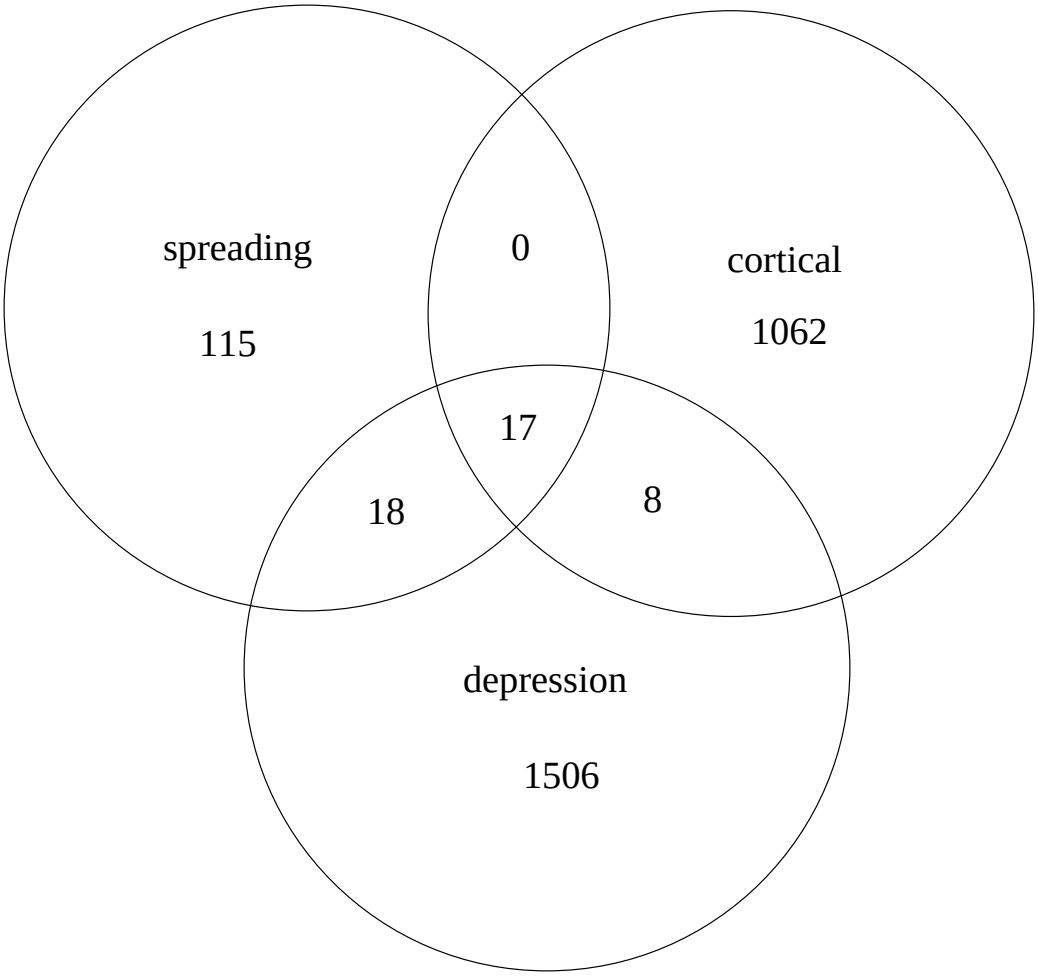


(A OR B) AND C



$$n(A \text{ OR } B) = n(A) + n(B) - n(A \text{ AND } B)$$

$$175 = 100 + 100 - 25$$


spreading
115
0
cortical
1062
17
18
8
depression
1506