Comparison of Markush Structure Databases

Hajime Tokuno

Intellectual Property Department, Mitsubishi Kasei Corporation, 5-2 Marunouchi 2-chome, Chiyoda-ku, Tokyo 100, Japan

Received March 10, 1993*

Search results from three Markush structure files, WPIM, MPHARM, and MARPAT, were compared using actual structure queries. Results from MARPAT were combined with those from Registry/CA searches because those two files were complementarily designed by the database producer. The differences in search results were due to various reasons such as search systems where the files are loaded, quality of data, and the accuracy of query structures. The policies for covering and indexing patent documents by each database affected the results the most. WPIM and MARPAT/Registry/CA combination gave comparative results, although MARPAT/Registry/CA usually did better. PHARMSEARCH was generally inferior because it covers patents from fewer countries than those covered by the other two services. WPIM was good in retrieving patents involving pharmaceutical compositions which contain the query substances. Registry and thus CA were good in retrieving structures which were described in patent examples and not specifically claimed, e.g., starting materials and intermediates.

INTRODUCTION

The introduction of Markush structure search systems on Markush-DARC and STN was a big breakthrough for patent searchers in pharmaceutical companies. The databases on these systems, WPIM and MPHARM on Markush-DARC and MARPAT on STN, seem to be very useful in answering patent queries with Markush or generic structures. WPIM is a structure file covering both Markush/generic and specific structures found in claims of patents covered by the World Patent Index (WPI). It is produced by Derwent Publications Ltd. and covers patents indexed since 1987. MPHARM is a structure file similar to WPIM, corresponds to Pharmsearch produced by INPI, and contains patents since 1987.[1] MAPRAT, a product of Chemical Abstracts Service (CAS), is a Markush structure and bibliographic file covering patents which are also covered by the CA file and which contain Markush structures in their claims and/or disclosures.[2-3] It covers patents issued since 1988.

Several papers have been published already which compare these systems and databases. Barnard[4] discussed the differences of search system design between DARC, MARPAT, and GENSAL. Schmuff[5] compared features of search software on STN and DARC. Cloutier[6] compared coverages of WPIM, MPHARM, and MARPAT for U.S. patents published on a certain date, as well as indexing of five documents. Since this study was conducted at a very early stage of database development, her observation seems to be very preliminary. Wilke[7] discussed the difficulty of searching for small organic compounds by any search systems including MARPAT and Markush-DARC. Fanzreb et al.[8] examined WPIM briefly in comparison with WPI, GREMAS, and CA. The WPIM database was too immature at the time of the study to give any meaningful results.

The Japan Farmdoc Association is a user group of 38 pharmaceutical companies which subscribe to the information service provided by Derwent. It began evaluating these Markush services as early as their beta testing stages. A study group was formed in its Eastern Section and has been making an extensive comparison of the services for several years. This paper is a summary of two studies, the first conducted from late 1990 to early 1991 and the second from late 1991 to early 1992. The purpose of the studies was to compare the databases—their scope, coverage, indexing policy, and reliability—by conducting actual searches. System features are not discussed unless they were critical in performing a particular search.

The MARPAT and Registry files were always treated as a single database in this comparison. This treatment is in accordance with our daily practice to search MARPAT always together with Registry, since these two files are complementary with each other. Markush (generic) structures described in claims and disclosures of patents are covered only by MARPAT. Specific structures described in claims, as well as those described in disclosure examples, are indexed only by CA, and at the same time registered in the Registry file.

The comparison of search results had some technical difficulties. There were several reasons for which a particular patent was not found in a database, other than system problems and encoding errors. First, the particular patent was out of the scope of the database, because the issuing country was not covered, or because it was not a basic patent for the database, etc. In particular, CA, and hence MARPAT, does not normally cover continuation-in-part patents of a particular patent, while WPIM does.

Second, the indexing policy of one database does not allow indexing of a specific structure. MARPAT does not index any specific structures. Claimed structures are, however, indexed by CA and thus would be found in Registry. This required us always to combined results from MARPAT with those from Registry/CA. Specific substances in patent examples are covered by WPIM as specific variations derived from Markush structures. They are also covered by Registry/CA but not by MARPAT. Again the combination of MARPAT and Registry/CA is necessary to make a thorough comparison.

Third, the query structure might not be created accurately to retrieve all the possible answers. When this was the case, the results were adjusted to reflect the query intention.

There can be two types of flaws in databases. One flaw is missing patents. Some patents were missing because the database was not current enough. Some were missing because they were just waiting to be processed, or backlogged. Another

* Abstract published in *Advance ACS Abstracts,* October 1, 1993.

flaw is indexing errors. A certain portion of a structure was not coded correctly or not coded at all.

PROCEDURES

The four structure databases, WPIM, MPHARM, MARPAT, and Registry, were searched by using selected queries which represent a wide variety of pharmaceutical structures. Generic structures which represent a wider range of specific substances were chosen for the queries. Possible differences in structure queries resulting from differences in system features were kept to a minimum. The structure queries were run on each system.

All comparisons were made on the basis of patent citations; i.e., all structure search results were crossed over to the corresponding citation files, except those from MARPAT. Answers from WPIM and MPHARM have compound numbers (CN) as identifiers. These were then crossed over to WPI and PHARMSEARCH, respectively, to get the patent references. Answers from MARPAT are patent references themselves. Answers from Registry have the Registry Numbers as identifiers, which were crossed over to CA to retrieve the patent citations. The results from MARPAT were always combined (i.e., ORed) with the results from the Registry file crossed over to the CA file.

In the first study, nine queries were selected to represent a wide variety of current pharmaceutical topics. Hit references were limited to those with the patent publication year of 1989. Since WPIM covers patents indexed since 1987 and MARPAT covers patents pubished since 1988, both files were assumed to be fairly complete by that time. The results were examined to avoid any misinterpretation from differences in country coverages and policies of identifying patent families, especially between WPI and CA.

In the second study, nine queries were selected as well to represent current pharmaceutical topics of interest. Some were similar to the ones used in the first study. WPIM, MARPAT, and Registry were all searched by using them. MPHARM was not searched in this study. The references obtained were limited to the patent publication years of 1989 and 1990. The rating scheme was revised to be more detailed than what was used in the first study, i.e. to represent (1) the difference in country coverages and the policies of patent family, (2) the difference in indexing policies, (3) the effect of structure query conventions where the scope of a query structure might be broader or narrower than the actual query requirement, (4) indexing errors, (5) search system problems, (6) the absence of data at all, and (7) other reasons.

RESULTS

First Study (For 1989 Patents). The queries used in the first study are shown in Figure 1. The actual query structures created for each system are shown for query 5 in Figure 2. Table I shows the results of the first study. The total hits column indicates the total count of patents retrieved by all databases. The hits from individual databases are listed in the other columns. The numbers in parentheses are recalls, in percentages, obtained by dividing the number of hits of individual databases by the total hits.

Table II shows the comparison of MARPAT/Registry and WPIM in terms of unique hits between them. For example, in the case of the query captoprils, MARPAT/Registry retrieved two hits which were not retrieved by WPIM, while WPIM retrieved seven hits which were not found by MARPAT/Registry. The last column lists unique hits obtained solely from MPHARM and not from either WPIM or MARPAT/Registry.

Generally speaking, the combined searches of MARPAT and Registry gave better results than those of WPIM or MPHARM; i.e., the results of MARPAT/Registry were better than those of WPIM, except for queries 1, 4, and 8 (Table I), and better than those of MPHARM in all queries. The combined MARPAT/Registry also gave more unique hits than those of WPIM. This partly resulted from the evaluation procedure in which the indexing policy of WPIM was not properly taken into account.

Both WPIM and MARPAT were not complete in covering the necessary patents at the time of the searches. This fact resulted in many missing patents in these databases in some searches. Since MPHARM has narrower country and subject coverages, the file did not give many unique hits, as shown in Table II.

Table III shows the hits from MARPAT and those from Registry, which were crossed over to CA to get patent references. Searches of the Registry file are always complementary because patents without Markush structures (containing specific structures only) are not included in the MARPAT file. On the other hand, MARPAT quite often gave additional hits which were not found by Registry. This proved the necessity of using MARPAT in addition to Registry when comprehensive results are needed on CAS ONLINE.

MARPAT and Registry were successful in retrieving intermediates and starting materials but not always successful in retrieving patents on pharmaceutical compositions. The generic search feature of MARPAT is considered superior to that of Markush-DARC, because the former allows system mapping of generic groups (nodes) to specific groups and vice versa.

Although not shown in the tables, there were more indexing errors found in the WPIM file than in MARPAT/Registry. And the currency of the data was better for MARPAT.

Comments about individual searches are as follows.

Query 1: Captoprils. The search on WPIM gave better results (17/19) than those of MARPAT/Registry (12/19). The latter missed patents on pharmaceutical compositions using captoprils. Some patents with peptide structures (WO 8910961, WO 8910369, and DE 3831936) were not in MARPAT, possibly because the structures were too big to process. Two of them (WO 8910369 and DE 3831936) were later added to the database, and one is waiting for the coming enhancement of the input system.

Query 2: Cephalosporins. The searches on MARPAT/Registry gave intermediates and starting materials as hits, but that on WPIM did not. Thus, recalls were better for MARPAT/Registry (25/29) than WPIM (20/29). On the other hand, many patents treating pharmaceutical compositions were not retrieved by MARPAT/Registry. Two structures of WPIM were incorrectly coded so that they were not retrieved here.

Query 3: 5HTs. The searches on MARPAT and Registry retrieved all the answers (18/18), while that of WPIM gave 72% of the answers (13/18). The currency of WPIM was poor. One *N*-alkyl derivative was not retrieved in WPIM because CHK was not included in the query.

Query 4: Oxicams. The searches on WPIM gave better results (15/18) than that of MARPAT/Registry (11/18), probably because many were patents on pharmaceutical compositions using oxicams. Neither MARPAT nor Registry normally covers specific components of such composition patents.

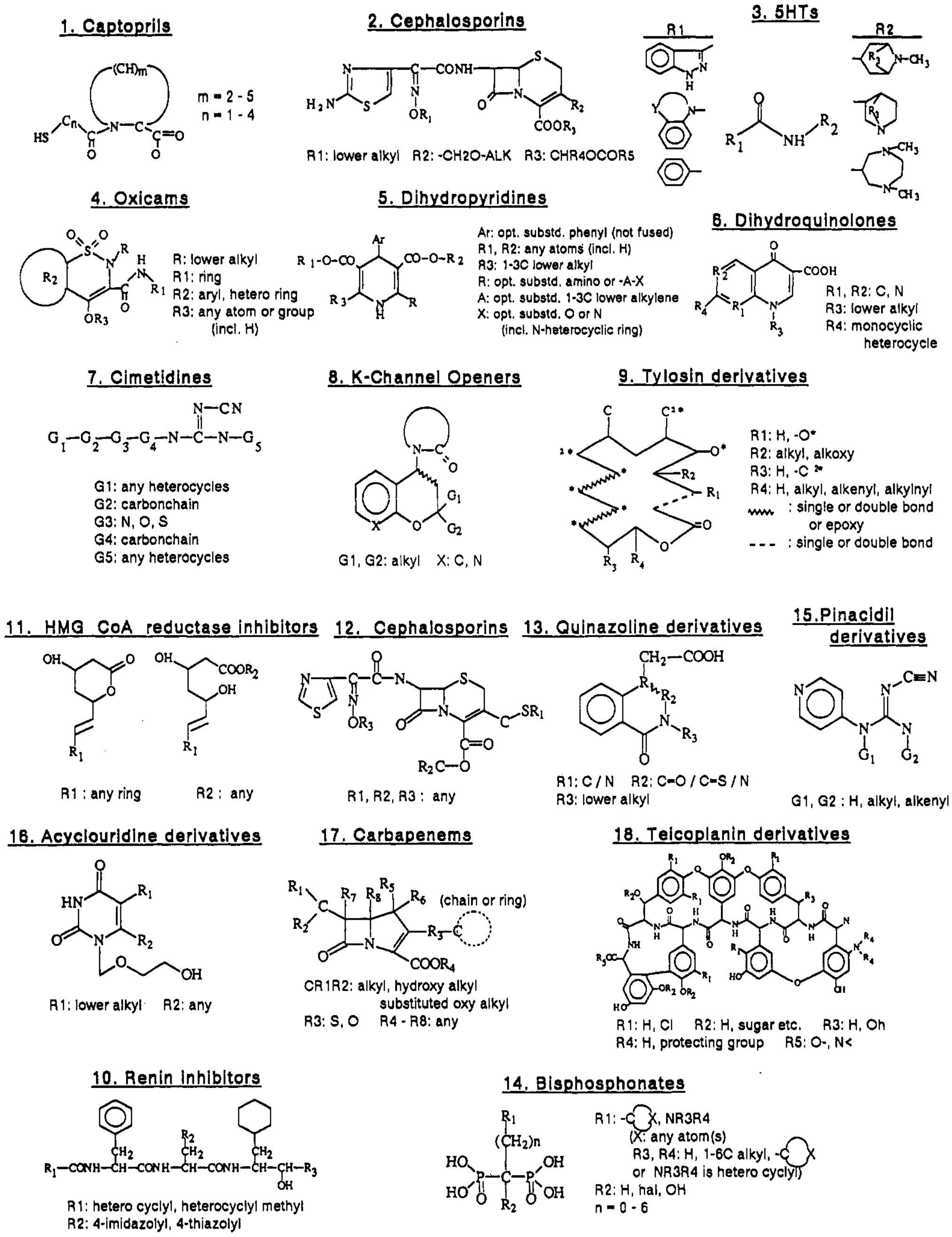

1. Captoprils
m = 2 - 5
n = 1 - 4
2. Cephalosporins
R1: lower alkyl R2: -CH2O-ALK R3: CHR4OCOR5
3. 5HTs
R1
R2
4. Oxicams
R: lower alkyl
R1: ring
R2: aryl, hetero ring
R3: any atom or group (incl. H)
5. Dihydropyridines
Ar: opt. substd. phenyl (not fused)
R1, R2: any atoms (incl. H)
R3: 1-3C lower alkyl
R: opt. substd. amino or -A-X
A: opt. substd. 1-3C lower alkylene
X: opt. substd. O or N (incl. N-heterocyclic ring)
6. Dihydroquinolones
R1, R2: C, N
R3: lower alkyl
R4: monocyclic heterocycle
7. Cimetidines
G1: any heterocycles
G2: carbonchain
G3: N, O, S
G4: carbonchain
G5: any heterocycles
8. K-Channel Openers
G1, G2: alkyl X: C, N
9. Tylosin derivatives
R1: H, -O*
R2: alkyl, alkoxy
R3: H, -C*
R4: H, alkyl, alkenyl, alkylnyl
: single or double bond or epoxy
: single or double bond
11. HMG CoA reductase inhibitors
R1 : any ring R2 : any
12. Cephalosporins
R1, R2, R3 : any
13. Quinazoline derivatives
R1: C / N R2: C=O / C=S / N
R3: lower alkyl
15. Pinacidil derivatives
G1, G2 : H, alkyl, alkenyl
16. Acyclouridine derivatives
R1: lower alkyl R2: any
17. Carbapenems
(chain or ring)
CR1R2: alkyl, hydroxy alkyl substituted oxy alkyl
R3: S, O R4 - R8: any
18. Teicoplanin derivatives
R1: H, Cl R2: H, sugar etc. R3: H, Oh
R4: H, protecting group R5: O-, N<
10. Renin inhibitors
R1: hetero cyclyl, heterocyclyl methyl
R2: 4-imidazolyl, 4-thiazolyl
14. Bisphosphonates
R1: NR3R4
(X: any atom(s)
R3, R4: H, 1-6C alkyl, or NR3R4 is hetero cyclyl)
R2: H, hal, OH
n = 0 - 6

Figure 1. Query structures: 1–9 are for the first study; 10–18 are for the second study.

Query 5: **Dihydropyridines.** The searches on MARPAT/Registry gave better results (18/21) than that on WPIM (14/21). There were four structures which contain errors in WPIM, but only one in MARPAT. WPIM missed four patents because of the delay of updating, while MARPAT missed five.

Query 6: **Dihydroquinolones.** The MARPAT/Registry combination gave a better result (34/35) than that of WPIM (29/35).

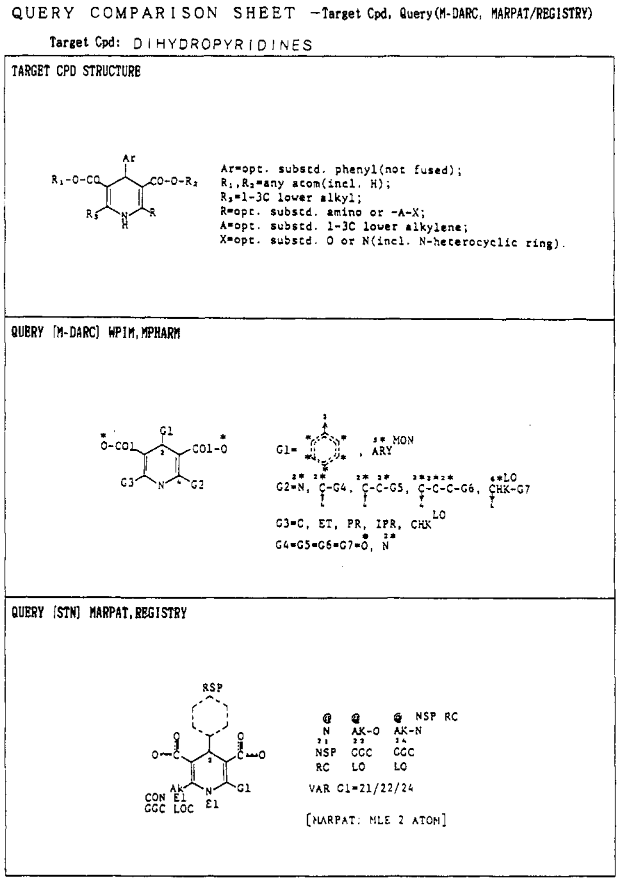
Query 7: **Cimetidines.** The MARPAT/Registry pair retrieved more hits (24/29) than WPIM (18/29). Many patents on pharmaceutical compositions claiming cimetidine itself as a component were not retrieved in Registry. These patents were not included in MARPAT because they did not contain Markush structures. The MARPAT/Registry pair gave a better result (six hits) in retrieving new compounds than WPIM (one hit). Markush-DARC had a problem in searching queries with generic requirements such as any heterocycles or any alkyls. Thus the query for WPIM tends to be narrower in scope than that of MARPAT, which resulted in favor of MARPAT.

Query 8: **K-Channel Openers.** The results from both WPIM and MARPAT/Registry were the same (14/15), and that from MPHARM was comparable (13/15). The one patent missed by WPIM was due to slowness in updating. The one missed by MARPAT/Registry was a composition patent. There was an error in one of the records in MARPAT.

Query 9: **Tylosin Derivatives.** The results from MARPAT/Registry (19/23) were slightly better than that from WPIM (18/23). It is unjustifiable that MARPAT did not cover some of the new compounds patented. Three Chinese patents (CN 1030791, CN 1031334, and CN 1034489) were not included in MARPAT.

Second Study (For 1989 and 1990 Patents). Queries used in the second study are shown in Figure 2. Tables IV–VI correspond to Tables I–III in the first study.

Out of nine queries, the MARPAT/Registry pair gave better recall than WPIM for six queries. The former gave recall of



QUERY COMPARISON SHEET —Target Cpd, Query(M-DARC, MARPAT/REGISTRY)

Target Cpd: DIHYDROPYRIDINES

TARGET CPD STRUCTURE

Ar=opt. substd. phenyl(not fused);
R_1,R_2=any atom(incl. H);
R_3=1-3C lower alkyl;
R=opt. substd. amino or -A-X;
A=opt. substd. 1-3C lower alkylene;
X=opt. substd. O or N(incl. N-heterocyclic ring).

QUERY [M-DARC] WPIM,MPHARM

G2=N, C-G4, C-C-G5, C-C-C-G6, CHK-G7

G3=C, ET, PR, IPR, CHK

G4=G5=G6=G7=O, N

QUERY [STN] MARPAT,REGISTRY

VAR G1=21/22/24

[MARPAT: MLE 2 ATOM]

Figure 2. Original query structure for query 5 (top), structure framed for WPIM and MPHARM (middle), and structure framed for MARPAT and Registry (bottom).

Table I. Comparison of Search Results from MARPAT/Registry, WPIM, and MPHARM in the First Study (Numbers in Parentheses Are Recalls in Percentage)

	total hits	MARPAT/ Registry	WPIM	MPHARM
1. captoprils	19	12 (63)	17 (89)	9 (47)
2. cephalosporins	29	25 (86)	20 (69)	10 (34)
3. 5HTs	18	18 (100)	13 (72)	13 (72)
4. oxicams	18	11 (61)	15 (83)	3 (17)
5. dihydropyridines	21	18 (85)	14 (67)	6 (29)
6. dihydroquinolones	35	34 (97)	29 (83)	11 (31)
7. cimetidines	29	24 (83)	18 (62)	12 (41)
8. *K*-channel openers	15	14 (93)	14 (93)	13 (87)
9. tylosin derivatives	23	19 (83)	18 (78)	14 (61)
total	207	175 (85)	158 (76)	91 (44)

Table II. Unique Hits between MARPAT/Registry and WPIM and Unique Hits from MPHARM in the First Study

	total hits	hits from MARPAT/ Registry but not from WPIM	hits from WPIM but not from MARPAT/ registry	hits from MPHARM only
1. captoprils	19	2	7	0
2. cephalosporins	29	8	3	1
3. 5HTs	18	5	0	0
4. oxicams	18	3	7	0
5. dihydropyridines	21	7	3	0
6. dihydroquinolones	35	6	1	0
7. cimetidines	29	9	3	2
8. *K*-channel openers	15	1	1	0
9. tylosin derivatives	23	4	3	1
total	207	45	28	4

Table III. Comparison of Search Results from MARPAT and Registry in the First Study (Numbers in Parentheses Are Those of Unique Hits between the Two Files)

	MARPAT + Registry	hits from MARPAT	hits from Registry
1. captoprils	12	3 (2)	10 (9)
2. cephalosporins	25	18 (2)	23 (7)
3. 5HTs	18	14 (0)	18 (4)
4. oxicams	11	4 (1)	10 (7)
5. dihydropyridines	18	9 (7)	11 (9)
6. dihydroquinolones	34	31 (8)	26 (3)
7. cimetidines	24	4 (2)	22 (20)
8. *K*-channel openers	14	10 (0)	14 (4)
9. tylosin derivatives	19	10 (0)	19 (9)
total	175	103 (22)	153 (72)

Table IV. Comparison of Search Results from MARPAT/Registry, WPIM, and MPHARM in the Second Study (Numbers in Parentheses Are Recalls in Percentage)

	total hits	MARPAT/ Registry	WPIM
10. peptide rennin inhibitors	21	18 (86)	9 (43)
11. HMG CoA reductase inhibitors	33	27 (81)	29 (88)
12. cephalosporins	29	24 (83)	19 (66)
13. quinazolines	13	11 (85)	11 (85)
14. bisphosphonates	24	20 (83)	18 (75)
15. pinacidils	14	13 (93)	9 (64)
16. acyclouridines	14	13 (93)	12 (86)
17. carbapenems	36	32 (89)	34 (94)
18. teicoplanins	15	15 (100)	11 (73)
total	199	173 (87)	152 (76)

over 80% for all queries, while the latter, for only four queries. Although the CAS files gave better results overall and again better currency, the currency, completeness, and quality of WPIM seem to have improved significantly since last year.

Since the rating scheme of this second study was more detailed than that of the first study, the results were tabulated in a different format in Table VII; i.e. results from the nine queries were combined together and classified, to highlight the characteristics of the databases. Contrary to the first study, we examined the irrelevant patents which were retrieved for some reason. Thus the totals of the hits of all searches in Table IV (and Table I in the first study) correspond to the line relevant patents in Table VII.

For retrieved patents, the MARPAT + Registry column shows the ORed results of hits of the two databases. For missed patents, the columns shows the ANDed results of the missed patents of the two databases. This AND operation can cause an overlap problem. When one patent was missed by MARPAT because of an indexing error, and it was also missed by Registry because of the inefficient query, then it was counted as a missed patent for the database combination. But, if the individual reasons (indexing error and query structure problem) were considered, ANDed results were hit rather than missed. Thus the MARPAT + Registry column was left open for all the relevant patents lines. The number 26 in the line (4) + (5) for MARPAT + Registry represents the independent count of relevant patents which was missed both from MARPAT and Registry. Since no overlap between the two reasons (indexing policy and coverage policy) of the out of scope category were found, their numbers were shown.

There were 199 records which satisfied the scope of the search queries out of 244 total patents retrieved. When individual databases were compared in terms of relevant patents retrieved, WPIM (152) and Registry (146) retrieved more patents than MARPAT (117). When, however, MARPAT and Registry were combined, they gave better performance in that the combination retrieved 173 hits (87%)

Table V. Unique Hits between MARPAT/Registry and WPIM in the Second Study

	total hits	hits from MARPAT/ Registry but not from WPIM	hits from WPIM but not from MARPAT/ Registry
10. peptide rennin inhibitors	21	12	3
11. HMG CoA reductase inhibitors	33	4	6
12. cephalosporis	29	10	5
13. quinazolines	13	2	2
14. bisphosphonates	24	6	4
15. pinacidils	14	5	1
16. acyclouridines	14	2	1
17. carbapenems	36	2	4
18. teicoplanins	15	4	0
total	199	47	26

Table VI. Comparison of Search Results from MARPAT and Registry in the Second Study (Numbers in Parentheses Are Those of Unique Hits between the Two Files)

	total hits	hits from MARPAT	hits from Registry
10. peptide rennin inhibitors	18	4 (2)	16 (14)
11. HMG CoA reductase inhibitors	27	24 (7)	20 (3)
12. cephalosporins	24	17 (8)	16 (7)
13. quinazolines	11	3 (2)	9 (8)
14. bisphosphonates	20	16 (0)	20 (4)
15. pinacidils	13	7 (5)	8 (6)
16. acyclouridines	13	10 (3)	10 (3)
17. carbapenems	32	24 (0)	32 (8)
18. teicoplanins	15	12 (0)	15 (3)
total	173	117 (27)	146 (56)

Table VII. Analysis of Search Results from MARPAT, Registry, and WPIM

	MARPAT	Registry	MARPAT + Registry	WPIM	total
retrieved patents (1)	150	156	210	161	244
relevant patents	117	146	173	152	199
irrelevant patents	33	10	38	9	45
query structure problem	9	1	9	3	
indexing error	9	8	14	4	
system problem	15	1	15	2	
missed patents (2)	94	88	34	83	
irrelevant patents (3)	12	35	7	36	
out of scope (4)	49	40	13	13	
indexing policy	38	28	2	12	
coverage policy	11	12	11	1	
relevant patents (5)	33	13		34	
query structure problem	8	8		6	
indexing error	11	3		15	
system problem	0	0		1	
records not indexed	14	2		12	
(4) + (5)	82	53	26[a]	47	
total (1) + (2)	244	244	244	244	

[a] Relevant patents missed by both MARPAT and Registry were counted independently.

compared to WPIM's 156 (78%). When the noises were examined, however, MARPAT/Registry was poorer, giving 38 irrelevant answers to WPIM's 9. The irrelevant answers resulted from either inappropriate query structures (9 from MARPAT/Registry and 3 from WPIM), indexing errors (14 from MARPAT/Registry and 4 from WPIM), and the limitation of the search system features (15 from MARPAT/Registry and 2 from WPIM). It is notable that MARPAT retrieved many irrelevant hits (15) due to the system limitation. This system limitation which caused incomplete iteration of queries was removed in mid-1992.

There were 94 unretrieved patents by MARPAT, 88 by Registry, and 83 by WPIM. Relevant patents which satisfied the scope of the corresponding queries were 82 of 94 for MARPAT, 53 of 88 for Registry, and 47 out of 83 for WPIM. The numbers of relevant patents which were missed by each database show WPIM was best, followed by Registry and then MARPAT. When MARPAT and Registry were combined, however, the number of relevant but missed patents by them were smaller (26) than that for WPIM (47).

However, 49 of the 82 for MARPAT, 40 of the 53 for Registry, and 13 of the 47 for WPIM had not been included in those databases because of their document coverage and indexing policies. Thus 33 for MARPAT, 13 for Registry, and 34 for WPIM were pure retrieval problems. When examined, 8 for MARPAT, 8 for Registry, and 6 for WPIM resulted from inappropriate structure queries, 11 for MARPAT, 3 for Registry, and 15 for WPIM were erroneously indexed, and 1 for WPIM resulted from a system problem which was later fixed. And, 14 for MARPAT, 2 for Registry, and 12 for WPIM were not included in the files at all. MARPAT was and still is not complete for 1989–90 patents, while WPIM was and is not complete in and before the 19th Derwent week of 1989.

Query 10: Peptide Rennin Inhibitors. WPIM gave poor results (8/20) because many (10) of the nonretrievals were not encoded in the first place, probably because the peptide structures were very large. This was also true for MARPAT in that eight corresponds had not been input yet. The combined results of MARPAT and REGITRY was good (18/20).

Query 11: HMG CoA Reductase Inhibitors. WPIM gave better results (29/33) than that of MARPAT/Registry (27/33).

Query 12: Cephalosporins. The searches on MARPAT and Registry retrieved 83% of all answeres (24/29), while WPIM gave only 66% (19/29). An encoding error was found in one record of WPIM. Three records were retrieved by Registry only because the corresponding structures were intermediates. The alkyl portion of the ester was not explicitly coded in four records of WPIM. They were coded by MARPAT. On the other hand, two structures had not been encoded in MARPAT.

Query 13: Quinazolines. The numbers of hits were the same (11/13) for both WPIM and the MARPAT/Registry pair. The missed patents were different, however. Two patents not retrieved by MARPAT/Registry had family patents which were covered by CAS in a different year range. Two patents which were not retrieved by WPIM had the hit structures as intermediates, which are not usually indexed according to its policy.

Query 14: Bisphosphonates. The searches on MARPAT/Registry gave slightly better results (20/24) than that on WPIM (18/24).

Query 15: Pinacidils. The search on MARPAT/Registry retrieved most (13/14) of the relevant answers while WPIM retrieved nine (9/14). One patent which was not retrieved was about pharmaceutical composition. However, the CA file indexed the query structure as a general derivative /D. This shows the importance of including such /D-substances in *L*-number crossover searching from the Registry file. Two structures were not retrieved by WPIM because the query structure was insufficient in that it did not contain HEA (heteroaromatics) in addition to a pyridine ring. Mapping to a broader group is handled automatically by MARPAT. One records had not been input yet in WPIM.

Query 16: **Acyclouridines.** The result from MARPAT/Registry (13/14) was slightly better than that from WPIM (12/14). WPIM had one structure error. Another patent was not retrieved by WPIM because the corresponding structure was out of the scope of its indexing policy, i.e., a starting material, which is not indexed although it was mentioned in the claim. A nonretrieval from MARPAT resulted because the examples of variable fragments were incomplete.

Query 17: **Carbapenems.** WPIM (34/36) gave better results than the MARPAT/Registry combination (32/36). Two patents were not retrieved from WPIM because their structures were incorrectly indexed. Of four patents which were not retrieved from MARPAT/Registry, one had structure errors, one was not yet indexed by CAS at that time, one was a composition patent whose substances mentioned were not necessarily indexed by either Registry or MARPAT, and one was out of the scope of MARPAT because it does not have any Markush structures and at the same time was indexed by Registry as a family patent which was out of the year range defined here.

Query 18: **Teicoplanins.** MARPAT/Registry gave all hits retrieved here (15/15), while WPIM gave 11 hits out of 15. Two patents were not retrieved in WPIM because of indexing errors, and one was not yet indexed at the time of the search.

MARPAT and Registry. When just the MARPAT and Registry files were compared, MARPAT retrieved unique hits in 12 cases out of 18 searches as shown in Tables III and VI. The complementary nature of the two databases was also demonstrated in Table VII. The number of patents which were not retrieved because they were out of the scope of the indexing policy of the corresponding databases were significantly larger for MARPAT (38) and Registry (28) than for WPIM (12). But when MARPAT and Registry were combined the number was only 2. This clearly shows that it is necessary to search MARPAT in addition to Registry to find chemical structure patents.

DISCUSSION

It is now very clear that the reliability of data and the capabilities of both search systems have been improved significantly. Various flaws in the systems and errors in the databases which we noticed when they were first introduced have mostly disappeared. The general observation was that the overall performance of the databases was in the order of MARPAT/Registry, WPIM, and then MPHARM. If one looks into individual queries, however, this order was sometimes reversed or equaled between MARPAT/Registry and WPIM. We conclude it is rather difficult to choose one system superior to another.

One of the major reasons which gave the above order was the differences in indexing policies, including country coverages and subject coverages. If those factors are compensated for, the difference in performance may become much closer. The current results are, however, very meaningful.

Another reason for the order was the differences in search capabilities. This sometimes resulted in differences in the scopes of queries. Our purpose, however, was to find out if a combination of the particular databases on a particular system gives us reasonable results. Thus the differences in system capabilities were basically ignored.

We successfully revealed the effects of indexing policy on search results, especially from the second study. Since the current differences in indexing policies will not disappear for some time, it is desirable for us to use all of these databases together.

We also discovered other characteristics of the studied databases. For example, WPIM indexed claimed pharmaceutical compositions better than the other databases. The Registry file, on the other hand, indexed more starting materials and intermediates than the other databases. The combination of MARPAT and Registry performed better for more generic structure queries, such as those containing ANY heterocycles, than WPIM and MPHARM. Although MPHARM covers fewer patent countries, it was the most current at the time of the first study.

As mentioned earlier, the systems were being enhanced and the data were being corrected even during our studies. We cannot reproduce the same results now. But we believe the aim, methods, results, and issues discussed in our study are still valid and useful information for those who search patent databases, as well as producers of databases and engineers of systems.

ACKNOWLEDGMENT

This study was the result of cooperative work by the following members of the Japan Farmdoc Association: Naoko Nakayama, Chikako Watanabe, Hitoshi Furuyama, Kimie Aiba, Nobuo Takano, Hiroyuki Hirano, Tadashi Kuribayashi, Kazuaki Takeuchi, Hitoshi Ikeda, Atsuo Matsunaga, Makoto Yoshida, Noriko Irokawa, Noriko Minamizaki, and Hiroko Tsuji. I wish to express deep appreciation for their cooperation. I also thank Soichi Tokizane and Isao Togo of the Japan Association for International Chemical Information, Toshiyuki Kaneko of Nippon Gijutsu Boeki Co., Ltd., Anthony Ferns of Derwent Publications Ltd., and Sumio Nagashima of Zenyaku Kogyo K.K. for their advise and help.

REFERENCES AND NOTES

(1) O'Hara, Michael, P.; Pagis, Catherine. The PHARMSEARCH Database. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 59–63.

(2) Fisanick, William. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145–154.

(3) Ebe, Tommy; Sanderson, Karen, A.; Wilson, Patricia, S. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 2. The MARPAT File. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 31-36.

(4) Barnard, John. M. A Comparison of Different Approaches to Markush Structure Handling. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 64–68.

(5) Schmuff, Norman, R. A Comparison of the MARPAT and Markush DARC Software. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 53–59.

(6) Cloutier, Kathleen A. A Comparison of Three Online Markush Databases. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 40–44.

(7) Wilke, Robert N. Searching for Simple Generic Structures. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 36–40.

(8) Franzreb, Karl Heinz; Hornbach, Pia; Pahde, Claudia; Ploss, Gottfried; Sander, Juergen. Structure Searches in Patent Literture: A Comparison Study between IDC GREMAS and Derwent Chemical Code. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 284–289.