REFERENCES AND NOTES

(1) Barnes, S. B., "The Beginnings of Learned Journalism", *Sci. Monthly*, 257–260 (March 1934).
(2) Becker, J., and Hayes, R. M., "Information Storage and Retrieval", Wiley, New York, 1963.
(3) Bernier, C. L., "Condensed Technical Literatures", *J. Chem. Doc.*, **8**, 195–197 (1968).
(4) Bolton, H. C., "A Catalogue of Scientific and Technical Periodicals, 1665–1895", Smithsonian Institute, Washington, D.C., 1897.
(5) Borko, H., and Chatman, S., "Criteria for Acceptable Abstracts: A Survey of Abstractors' Instructions", *Am. Doc.*, **14**, 149–160 (1963).
(6) Borko, H., and Bernier, C. L., "Abstracting Concepts and Methods", Academic Press, New York, 1975.
(7) Brown, C. H., "Scientific Serials", Association of College and Reference Libraries, Chicago, 1956.
(8) Collison, R., "Abstracts and Abstracting Services", Clio, Santa Barbara, Calif., 1971.
(9) Crane, E. J., Patterson, A. M., and Marr, E. B., "A Guide to the Literature of Chemistry", Wiley, New York, 1957.
(10) Durant, W., and Durant, A., "The Story of Civilization", 9 vols., Simon & Schuster, New York, 1954–1965.
(11) Hirayama, K., "Status of Chemical Information Activities in Japan", *J. Chem. Doc.*, **13**, 21–23 (1973).
(12) Keenan, S., and Elliott, M., "World Inventory of Abstracting and Indexing Services", *Spec. Libr.*, **64**, 145–150 (1973).
(13) Kronick, D. A., "A History of Scientific and Technical Periodicals (1665–1790)", Scarecrow Press, Metuchen, N.J., 1961.
(14) Leicester, H. M., "The Historical Background of Chemistry", Wiley, New York, 1956.
(15) Luhn, H. P., "The Automatic Creation of Literature Abstracts (Auto-abstracts)", *IBM J. Res. Dev.*, **2**, 159–165 (1958).
(16) Manzer, B. M., "The Abstract Journal, 1790–1920", Scarecrow Press, Metuchen, N.J., 1977.
(17) Osborn, A. D., "Serial Publication", American Library Association, Chicago, 1973.
(18) Partington, J. R., "A History of Chemistry", Macmillan, London, 1962.
(19) Skolnik, H., "The Multiterm Index: A New Concept in Information Storage and Retrieval", *J. Chem. Doc.*, **10**, 81–85 (1970).
(20) Skolnik, H., "Milestones in Chemical Information Science", *J. Chem. Inf. Comput. Sci.*, **16**, 187–193 (1976).
(21) Skolnik, H., "A Computerized Current Awareness System for Journal Literature", *J. Chem. Inf. Comput. Sci.*, **17**, 75–78 (1977).
(22) Skolnik, H., "Chemistry Two Centuries Ago", *J. Chem. Inf. Comput. Sci.*, **18**, 2A (1978).
(23) Weil, B. H., "Standards for Writing Abstracts", *J. Am. Soc. Inf. Sci.*, **21**, 351–357 (1970).
(24) Williams, T. I., "Communication in Science", *Chem. Ind. (London)*, 326–332 (May 7, 1977).
(25) Witty, F. J., "The Beginnings of Indexing and Abstracting. Some Notes Toward a History of Indexing in Antiquity and the Middle Ages", *Indexer*, **8**, 193–198 (1973).

Using Bibliometric Analyses of Patent Literature for Predicting the Clinical Fates of Developing Drugs†

DONALD A. WINDSOR

Norwich-Eaton Pharmaceuticals, P.O. Box 191, Norwich, New York 13815, and School of Advanced Technology, State University of New York, Binghamton, New York 13901

Received June 1, 1979

Certain bibliometric features of the early literatures of developing drugs can be used to predict their ultimate clinical fates. The chronological sequence of publications is expressed as a binary vector with 1 for a patent and 0 for a nonpatent. The decimal equivalents for standardized vector lengths provide scalar values for comparing one drug with another. In order to incorporate concordant patents, fuzzy subsets are employed, with the number of attempts required to achieve transitive closure being the values for comparison. The methods involved are described using minoxidil as an example.

"Just as the twig is bent the tree's inclined."[1] This may well be true for certain biological situations, but there is some indication to suspect that it may also hold for literature-based information systems. Information tends to spring from previous information, knowledge seems to beget more knowledge, and publications emerge from previous papers. At least, it was felt, such an approach to the literature might be worth investigating.

Consider the publications on a drug as forming a discrete information system. A drug literature does not spontaneously happen. Each publication is the end result of a dynamic interaction involving authors, editors, referees, printers, publishers, and even subscribers and readers. Once published, the paper becomes an individual element in another complex system, the body of literature itself. Unlike a static filing operation, a published document is related to other documents by the citation process.[2] The literature is a dynamic system composed of a citation network which changes each time a new publication cites previous papers.

The human forces that launch a publication all have their own motives. What they are is not too important, but their existence, per se, can be exploited.

† Presented at Annual Meeting, Science Information Subsection, Pharmaceutical Manufacturers Association, March 6, 1979.

As documents are published, they form numerous patterns with respect to one another. One paper comes before the next; one is a patent and the other is not; one reports on humans while another is about rats while still another is about chromatography; several appear in the same journal whereas others appear elsewhere, and so on.

These publication patterns are the combined outward expressions of all the social motives that launch the publications. As such, these patterns are characteristic of their antecedent literatures, fingerprints of individuality as it were—reflections of the motives that generated them. Or, put more simply, "by their fruits you will know them".[3]

When it comes to new drugs, the people who know the most about a particular drug are those who work with it. Investigators who realize they have a great drug will act with motives far different than if they have a mediocre drug. The resulting publication patterns reflect this. Therefore, if these patterns can be recognized, classified, and compared, they can be used to discern the great drugs from the mediocre ones, at a very early stage in their development—which is, in effect, a prediction of clinical success.

Patents are publications.[4] They must be treated as such. No bibliography on a drug is complete if it omits the patents. Unfortunately, the regular journal literature does not often cite the patent literature, and vice versa! There is a curious

0095-2338/79/1619-0218$01.00/0

Table I. Chronological Listing of the First 14 Publications on Minoxidil

MINOXIDIL CHRONOLOGICAL LISTING

1 UPJOHN CO. <SYNTHESIS OF 1,2-DIHYDRO-1-HYDROXY-PYRIMIDINES.> NETHERLANDS PATENT APPL. 6,615,385. 68 P. (MAY 2) 1967

0 MONTGOMERY, R. G. AND DUCHARME, D. W. HEMODYNAMIC EFFECTS OF CHRONIC AND ACUTE CHANGES IN BLOOD PRESSURE. FED. PROC. 27(2): 758, #3043 (MARCH-APRIL) 1968

0 ANON. U-10858. UNLISTED DRUGS 20(4): 520 (APRIL) 1968

1 ANTHONY, W. C. AND URSPRUNG, J. J. (UPJOHN) 6-AMINO-1,2-DIHYDRO-1-HYDROXY-2-IMINO-4-PHENOXYPRIMIDINES. U. S. PATENT 3,382,247. ?? P. (MAY 7) 1968

1 ANTHONY, W. C. AND URSPRUNG, J. J. (UPJOHN) 6-AMINO-4,5-DI(SUBSTITUTED AMINO)-1,2-DIHYDRO-1-HYDROXY-2-IMINOPYRIMIDINES. U. S. PATENT 3,382,248. ?? P. (MAY 7) 1968

0 CHIDSEY, C. A.; GILMORE, E. B.; AND WEIL, J. V. CLINICAL STUDIES WITH A NEW ANTIHYPERTENSIVE VASODILATOR, 6-AMINO-1,2-DIHYDRO-1-HYDROXY-2-IMINO-4-PIPERIDINO-PYRIMIDINE (POP). INT. CONGR. PHARMACOL., 4TH, (BASEL) ABSTR.: P. 308 (JULY) 1969

0 DUCHARME, D. W.; CHREST, C. C., AND MONTGOMERY, R. G. COMPENSATORY HEMODYNAMIC CHANGES ASSOCIATED WITH CHRONIC VASODILATION IN DOGS. INT. CONGR. PHARMACOL. 4TH (BASEL) ABSTR.: P. 308-309 (JULY) 1969

0 ANON. POP (U 10858). UNLISTED DRUGS 21(10): 157I (OCT.) 1969

0 GILMORE, E.; WEIL, J., AND CHIDSEY,C. TREATMENT OF ESSENTIAL HYPERTENSION WITH A NEW VASODILATOR IN COMBINATION WITH BETA-ADRENERGIC BLOCKADE. NEW ENGL. J. MED. 282(10): 521-527 (MARCH 5) 1970

0 USAN COUNCIL ON DRUGS. LIST NO. 92; NEW NAMES ... MINOXIDIL. JAMA 213(8): 1325-1326 (AUG. 24) 1970

0 ANON. MINOXIDIL. UNLISTED DRUGS 22(11): 1670 (NOV.) 1970

0 GOTTLIEB, T. AND CHIDSEY, C. EVALUATION OF RELATIVE THERAPEUTIC EFFICACY AND PHARMACOKINETICS OF MINOXIDIL, A NEW ANTIHYPERTENSIVE. CLIN. RES. 19(2): 349 ( ) 1971

1 ANTHONY, W. C. (UPJOHN CO.) 1,2-DIHYDRO-1-HYDROXY-2-IMINOPYRIMIDINES. GERMAN PATENT OFFEN. 2,114,887. 19 P. (OCT. 14) 1971

0 GOTTLIEB, T. B.; KATZ, F. H., AND CHIDSEY, C. A. COMBINED THERAPY WITH VASODILATOR DRUGS AND BETA-ADRENERGIC BLOCKADE IN HYPERTENSION. A COMPARATIVE STUDY OF MINOXIDIL AND HYDRALAZINE. CIRCULATION 45(3): 571-582 (MARCH) 1972

twist to this situation. The papers which would be the most apt to cite patents (i.e., the early papers) are very often published before the patents are. Most bibliographies on drugs will show this when the citations are arranged in a strictly chronological sequence. It is very unfortunate that *Chemical Abstracts* does NOT give the actual dates in its references, just the years. However, when the papers are inspected and the actual dates of publication are obtained (i.e., month and day), the chronological relationship of patents to journal articles is quite striking.

Table II. Country, Sequence, and Type Values for Patents[a]

country		sequence	type
Belgium	= 0.4	1st = 1.000	main = 1.0
Britain	= 0.8	2nd = 0.500	concordant = 0.5
Canada	= 0.8	3rd = 0.250	
France	= 0.7	4th = 0.125	
Germany East	= 0.6	5th = 0.063	
Germany West	= 0.8	6th = 0.032	
Japan	= 0.6	7th = 0.016	
Netherlands	= 0.7	8th = 0.008	
South Africa	= 0.5	9th = 0.004	
Switzerland	= 0.4	10th = 0.002	
United States	= 1.0	11th = 0.001	
U.S.S.R.	= 0.1	12th = 0.000	

[a] The membership grade is equal to their product.

Consider a bibliography on a specific drug in which all publications on this drug have been ranked chronologically from the earliest paper to the most recent. Assign the value 1 to each citation if it is a patent and 0 if it is not. The result will be a binary membership vector. For example, Table I is a bibliography which cites the first 14 papers on minoxidil. Its binary vector is listed in the left margin.

This vector is a characteristic fingerprint of the patent/nonpatent literature of minoxidil. The more elements (i.e., papers) in this vector, the higher the probability of its being unique. For example, the first paper by itself gives a vector of 1, hardly unique; the first two papers give 1 0; the first three give 1 0 0. The number of possible different arrangements of a binary vector is 2^n, where n is the number of papers.[5] For example, there are $2^5 = 32$ possible binary vectors for five papers, from 0 0 0 0 0 through 1 1 1 1 1.

The great advantage of using binary vectors is the ease of making comparisons. One vector is just matched, on an element-by-element basis, to another vector. The result can even be another binary vector, if desired, by just assigning a value of 1 whenever both elements are the same. For example, 1 0 0 1 1 0 compared with 1 0 0 1 0 1 is 1 1 1 1 0 0. An additional benefit is that a binary vector can be translated into its decimal equivalent.[6] For example, 1 0 0 1 1 0 is 38 and 1 0 0 1 0 1 is 37. As a single number, it can now easily be used as a domain element in a function where the range can be such real world entities as sales, patients treated, incidence of adverse reactions, etc. Consider, for example, comparing four hypothetical drugs, A, B, C, and D, to sales-dollars-per-year. If these drugs had six-element binary vectors of A = 0 0 1 0 1 0, B = 1 0 0 1 0 1, C = 0 1 0 0 0 0, and D = 0 0 0 0 0 1, their decimal representations would be $A = 10$, $B = 37$, $C = 16$, and $D = 1$. If the annual sales in millions of dollars were: $A = 3$, $B = 0.5$, $C = 1$, and $D = 5$, then an obvious relation would obtain where the higher the decimal equivalent of the binary vector, the lower the annual sales revenue. It is noticeable that the sooner a patent is published (in relation to other publications), the lower the sales figure. This was only a hypothetical example. The actual relationship, if there indeed is any, must, of course, be determined with real data.

The relationship of patents to other publications is more involved than the binary model indicates. Another bibliometric element of immense importance is the consideration of "human-papers". A human-paper is a paper that reports the administration of the drug to a human.[7] In order to incorporate them into the patent scheme, a ternary vector is needed, using 0, 1, and 2. A value of 2 is assigned to any human-paper, 1 to a patent, and 0 to every other kind of publication. Using the minoxidil example in Table I, the ternary vector is 1 0 0

Table III. Fuzzy Membership Matrix for Minoxidil

	Belg	Brit	Can	Fr	Ger E	Ger W	Jap	Neth	S Afr	Swiss	US	USSR
Belg	0	0	0	0	0	0	0	0	0	0	0	0
Brit	0	0	0	0	0	0	0	0	0	0	0	0
Can	0	0	0	0	0	0	0	0	0	0	0	0
Fr	0	0	0	0	0	0	0	0	0	0	0	0
Ger E	0	0	0	0	0	0	0	0	0	0	0	0
Ger W	0	0	0	0.09	0	0.40	0	0	0	0	0.13	0
Jap	0	0	0	0	0	0	0	0	0	0	0	0
Neth	0	0.30	0	0.26	0	0.39	0	0.70	0	0	0.47	0
S Afr	0	0	0	0	0	0	0	0	0	0	0	0
Swiss	0	0	0	0	0	0	0	0	0	0	0	0
US	0.03	0.05	0	0.04	0	0.05	0.04	0.04	0.03	0	0.25	0
USSR	0	0	0	0	0	0	0	0	0	0	0	0

Table IV. Algorithm for Transitive Closure

```
matrix(1)
counter = counter + 1
matrix(2) = matrix(1) + matrix(1)
matrix(3) = matrix(1) + matrix(2)
If matrix(3) ≠ matrix(1)
  then transitive closure has not
    been achieved.
    do again
If matrix(3) = matrix(1)
  then transitive closure has been achieved
    in counter attempts.
```

1 1 2 0 0 2 0 0 2 1 2 with its decimal equivalent of 1 686 686. The decimal equivalent of the binary vector is 9730. Binary vectors must be compared to other binary vectors, and ternary vectors to other ternary vectors; they cannot be mixed.

Other qualities of drug papers can, of course, be incorporated, such as nomenclature papers, or double-blind studies, or adverse reaction papers. The vector can be any-number-ary (*n*-ary), depending on the particular intent.

The bibliography in Table I suffers from an important omission. It does not contain the "equivalent" patents! The first paper published on minoxidil (as far as I can determine) is Netherlands Patent 6 615 385 on May 2, 1967. *Chemical Abstracts* cites 14 equivalent patents in its 8th and 9th Collective Index Patent Concordances. These patents are given by country and by number. That is all!

While there are indeed ways to get the necessary bibliometric data, such as dates, these ways are not accessible in small research or academic libraries. Therefore, other approaches have been tried. One of these bears mentioning here because it attempts to use the equivalent patents with only their countries needed. These patents will be referred to as "concordant" because the term "equivalent" has a very specific, and quite different, meaning in set theory.[8]

Chemical Abstracts prints an abstract for the first patent it receives on any given invention. No abstract is printed for subsequent patents on the same invention, but instead they are cross-referenced in the Patent Concordance by country and number.[9] There are 27 countries listed in the Introduction to the 9th Collective Index Patent Concordance. The main patents can be ranked sequentially by date, since the dates are given. The concordant patents do not have their dates given; however, they can be ranked sequentially within each country by number. Therefore, from this aspect there are three qualities of patents: type (main or concordant), country, and sequence.

When considering the literature in terms of binary vectors, the membership criterion was simply: was the document a patent or not? Even the ternary vector, although providing two questions (patent? human?) still dealt with a yes or no answer to both. However, when expanding into the realm of the concordant patents, the answers are no longer just yes or no. For example, Netherlands Patent 6 615 385 has two British, two French, six German, and four United States concordant patents. Yes, this patent is covered in France and in Germany, but is three times as much covered in Germany as in France meaningful? No, of course not. The patent has merely been divided into separate patents, but without looking at the actual patent it is impossible to tell which concordant patents contain the drug of interest.

Rather than operating with such a system in simple yes or no terms, it might be advantageous to confront it from a fuzzy set approach. Fuzzy reasoning assigns a membership value in the interval 0 through 1.[10] That is, rather than yes or no, it allows for grades of membership between yes or no. Since a fuzzy set must be a subset of a well-defined set, a relation on a set was established involving the set of all countries that issue patents. However, this seemed impractical because many of these countries would not be involved. Therefore, the set was arbitrarily limited to the 12 main countries listed by the Derwent FARMDOC.[11] Interest values were then assigned to each country. Sequence values were assigned by starting at unity and halving each sequence step. Types of patent values were assigned by giving 1.0 to a main patent and 0.5 to a concordant. The overall membership grade was obtained by multiplying the country value by the sequence value by the type value. These values are listed in Table II.

A fuzzy membership matrix can then be constructed for each drug. The main patent is mapped from its country to each of its concordant countries with its calculated membership grade. For example, a United States main patent that is the

Table V. Transitive Closure Matrix of the Minoxidil Matrix in Table III

	Belg	Brit	Can	Fr	Ger E	Ger W	Jap	Neth	S Afr	Swiss	US	USSR
Belg	0	0	0	0	0	0	0	0	0	0	0	0
Brit	0	0	0	0	0	0	0	0	0	0	0	0
Can	0	0	0	0	0	0	0	0	0	0	0	0
Fr	0	0	0	0	0	0	0	0	0	0	0	0
Ger E	0	0	0	0	0	0	0	0	0	0	0	0
Ger W	0.03	0.05	0	0.09	0	0.40	0.04	0.04	0.03	0	0.13	0
Jap	0	0	0	0	0	0	0	0	0	0	0	0
Neth	0.03	0.30	0	0.26	0	0.39	0.04	0.70	0.03	0	0.47	0
S Afr	0	0	0	0	0	0	0	0	0	0	0	0
Swiss	0	0	0	0	0	0	0	0	0	0	0	0
US	0.03	0.05	0	0.05	0	0.05	0.04	0.04	0.03	0	0.25	0
USSR	0	0	0	0	0	0	0	0	0	0	0	0

first patent published has a membership grade of 1.0 × 1.0 × 1.0 = 1.0 mapped to itself and a grade of 0.7 × 0.5 × 0.5 = 0.175 mapped to France. A West German patent that was the third patent has a membership grade of 0.8 × 0.250 × 1.0 = 0.2 to itself; however, if it were concordant rather than main it would have a grade of 0.8 × 0.250 × 0.5 = 0.1 from its main. The intent is to assign the highest weight to the first main and diminish the subsequent values. An example of the fuzzy matrix for minoxidil appears in Table III. In the 8th and 9th Collective Indexes, *Chemical Abstracts* cites three main patents for minoxidil: Netherlands 6 615 385 (May 2, 1967), West German 2 114 887 (Oct 14, 1971), and United States 3 910 928 (Oct 7, 1975).

The matrices for different drugs can be compared on a one-to-one basis. However, a search was undertaken to find some automatic means of making comparisons and arriving at conclusions.

The concept of the transitive closure[12] seemed attractive and was explored. A transitive relation can be defined as a relation from one element to a second element, and from the second element to a third, when the relation also holds from the first to the third, that is, (A,B), (B,C), (A,C). Transitive closure is obtained for a pair of relations (A,B) and (B,C) by effecting the relation (A,C). This closure is obtained by multiplying a membership matrix by itself and then taking the union of the inner product with the original matrix, and performing this operation until a constant result is achieved. The fundamental algorithm is stated in Table IV.

The operator * in that algorithm is regular matrix multiplication (inner product) for well-defined sets. For fuzzy sets, however, it can take several forms and many of the possibilities were investigated. The standard, so-called "max–min transitivity"[13] will be presented here. This means that each row element of one matrix (M_{ij}) is compared with its appropriate column counterpart of the other (M_{ji}), and the minimum value is taken. The results of each row x column cross are compared and the maximum value is taken as the element of the inner product matrix. That is, the maximum of the minima becomes the new element. In APL terminology this operator is tripartite and is $\lceil \cdot \lfloor$.[14] In BASIC the operator does not exist as such, but can be effected by a nested three-loop arrangement (e.g., loops I, K, J). The union is simply the maximum of each element (M_{ij}) comparison. In practical terms, it was advantageous to write programs in APL and then redo them in BASIC. The counter gives the number of times the operation had to be executed to achieve transitive closure. The minoxidil matrix in Table III gave the closure matrix in Table V in two attempts. Other drugs have given other results, so comparisons can be made; for example, baclofen achieved closure in only one attempt. The number of closure attempts, like the decimal expressions of the binary vectors, can be plotted against any other quantifiable aspects of clinical success.

Undoubtedly, the most important aspect of this kind of research is whether or not it works. Can bibliometrics be used to predict the clinical fates of developing drugs? The answer is "yes". Have I demonstrated this here? The answer is "no". Unlike a straight scientific subject, this one is heavily shrouded with proprietary considerations. Therefore, it is better to discuss methods, approaches, points of view, and even speculations, but leave the specific predictions in the private domain.

REFERENCES AND NOTES

(1) Pope, A. Epistle I, 1733. In: Bartlett, J. "Familiar Quotations", 13th ed.; Little, Brown and Co.: Boston, Mass., 1955, p 314a.
(2) Garfield, E. "Citation Indexing for Studying Science." *Nature (London)* **227**, 669–671 (1970).
(3) Matthew, St., Gospel, Chapter 7, Verse 20. In: "Holy Bible. The New Testament, Confraternity Edition"; Catholic Book Publishing Co.: New York, 1951, p 16.
(4) Maynard, J. T. "*Chemical Abstracts* as a Patent Reference Tool". *J. Chem. Inf. Comput. Sci.*, **17**, 136–139 (1977).
(5) Kemeny, J. G.; Snell, J. L.; Thompson, G. L. "Sets and Counting Problems. . . a General Principle". In: "Introduction to Finite Mathematics", 3rd ed.; Prentice-Hall: Englewood Cliffs, N.J., 1974; p 60.
(6) Smith, K. J. "Other Numeration Systems". In: "The Nature of Modern Mathematics", 2nd ed.; Brooks/Cole, Monterey, Calif., 1976; pp 98–104.
(7) Windsor, D. A. "Publications on a Drug before the First Report of Its Administration to Man". *Bull. Med. Libr. Assoc.*, **59**, 433–437 (1971).
(8) Preparata, F. P.; Yeh, R. T. "The Relations of Compatibility and Equivalence". In: "Introduction to Discrete Structures for Computer Science and Engineering"; Addison-Wesley: Reading, Mass., 1973; pp 39–47.
(9) Anon. "Introduction. Patent Concordance". *Chem. Abstr., 9th Collect. Index*, **76–85**, 1I (1972–1976).
(10) Ragade, R. K. "Fuzzy Sets in Communications Systems and in Consensus Formation Systems". *J. Cybernet.*, **6**, 21–38 (1976).
(11) Anon. "Country Coverage". Derwent Central Patents Index, Country Allerting Bulletin Sect B; FARMDOC 1978: Appendix I (Dec 13, 1978).
(12) O'Neil, P. E. and O'Neil, E. J. "A Fast Expected Time Algorithm for Boolean Matrix Multiplication and Transitive Closure". *Inf. Control*, **22**, 132–138 (1973).
(13) Klir, G. J.; Cavallo, R. "Binary Relations on a Set". In: "Fundamental Structures for Systems Science", in preparation.
(14) Gilman, L.; Rose, A. J. "Generalized Inner Product". In: "APL; an Interactive Approach", 2nd ed.; Wiley, New York, 1976; pp 240–248.