Reprinted from FEDERATION PROCEEDINGS
Vol. 16, No. 3, September, 1957
*Printed in U.S.A.*

# A UNIQUE SYSTEM FOR RAPID ACCESS TO LARGE VOLUMES OF PHARMACOLOGICAL DATA; APPLICATION TO PUBLISHED LITERATURE ON CHLORPROMAZINE

HARRIET E. ROCKWELL, ROBERT L. HAYNE AND EUGENE GARFIELD

# A UNIQUE SYSTEM FOR RAPID ACCESS TO LARGE VOLUMES OF PHARMACOLOGICAL DATA; APPLICATION TO PUBLISHED LITERATURE ON CHLORPROMAZINE

HARRIET E. ROCKWELL, ROBERT L. HAYNE AND EUGENE GARFIELD[1]

*From the Science Information Department, Smith, Kline and French Laboratories, Philadelphia, Pennsylvania*

IT HAS become increasingly evident in recent years that one of the major problems in the efficient use of scientific manpower is the need for easy and complete access to reports of work already done. Recognition of this need has led to the growth of documentation as a special field of science. Special techniques have had to be developed for the storage of large volumes of information and the rapid retrieval of this information when it is required by scientists. We would like to describe one of the techniques which the Science Information Department of Smith, Kline and French Laboratories has developed to handle pharmacological information.

Anyone who has worked with a new drug knows that there is usually a frustrating lack of information about it. When Smith, Kline and French became interested in chlorpromazine, we were faced with an unusual problem. We were presented with a large volume of information about the drug, which grew larger every week and soon threatened to swamp our facilities for keeping track of it. Long before we were ready to market Thorazine, it became apparent that we would have to devote our whole Science Information Department to this one drug if we did not find better ways of handling the data on it.

There are a number of well-known ways of handling an accumulation of reports and reprints.

*1*) They can be filed according to subject, either alphabetically or according to a numerical coding system. This works only with a small collection and only if authors have cooperated by discussing just one subject in each report.

*2*) They can be filed by date of receipt, accession number, author's name, or some other neutral identifier. In this case, a separate card index is often used to supplement the report file. This index can be as specific or as broad as the owner chooses. Card indexing will handle a considerably larger collection than subject filing of the original reports, and allows more thorough indexing of each document. It becomes unmanageable when the total number of index entries desired is high. In addition, it cannot easily identify a document that discusses more than one subject.

*3*) When a card index, with separate cards for each subject in each document, becomes too large for efficient use, some sort of a unit record is necessary. A unit record system involves a recording medium, such as punched cards, which still permits multiple indexing but requires only one card for each document. Punched cards can be searched either manually, using edge-punched cards, or by machine. Manual searching is inexpensive and is efficient for files up to about a thousand documents, depending upon the complexity of the searches required. Beyond this, machines become necessary. A primary characteristic of punched card systems is that all subject headings for a given document can usually be entered on one card; thus the card file contains only as many cards as there are documents. This obviously results in a considerable saving of filing space and time compared to a conventional card index. To be sure, in such a system, every card must be examined during each search. This basic disadvantage is offset, particularly in machine searching, by the speed of the machine and its ability to select simultaneously the cards which meet a complex set of characteristics.

The flood of information on chlorpromazine forced us to re-examine our methods for handling such data. Simple subject filing obviously could not handle several thousand detailed documents. A card index was a useful stop-gap, but it could not cope with the tide for very long. More important, it could not be used for correlating and tabulating data. An estimate of the amount of data we might expect to accumulate within a few years was enough to convince us that machine-sorted punched cards were the only way to keep the situation under control.

The wisdom of this decision soon became apparent. We now have more than 10,000 documents, published and unpublished, in our chlor-

[1] Present address: Woodbury, New Jersey.

promazine file. These documents can be divided into clinical and pharmacological reports, although there is considerable overlap. Since the needs of the clinicians and the pharmacologists for recall of data differ, these two types of information are handled somewhat differently, and only the pharmacological system will be discussed here. It should be pointed out, however, that all the data, clinical or pharmacological, in any one document are entered on one card, so that correlation between the two can be made if desired.

We think the system we have worked out for storing and retrieving pharmacological information is quite simple. It has to be, since machine documentation is only one of the activities of our busy Science Information Department. As with any punched card system, it is based on three essential factors: *1*) an index, or set of subject headings; *2*) a means of transferring this index to a pattern of holes in a punched card; and *3*) a machine which will examine these punched cards and sort out the ones which contain the desired subject headings. Each of these factors is discussed in detail below.

Anyone who has ever tried to index a file of reprints or abstracts knows that the first essential is a list of subject headings. The larger the file, the more necessary it is to have a comprehensive list that, as far as possible, anticipates all questions that might some day be asked of the file. A hundred documents can be re-indexed if the collector's point of view changes; 500 make the job pretty difficult, and 1000 make it impossible. We anticipated that the chlorpromazine pharmacology file might eventually grow to several thousand documents. Therefore, our first job was to set up a subject heading list or index that would include every concept we might ever need to use to describe their contents. The list should be specific enough to take care of all the chlorpromazine data, but it should also be general enough to apply to any drug. We believe our solution is unique and, at the same time, very simple.

The essential part of our subject heading list is a group of about 150 words or terms referring to organs, tissues or functions, and hormonal activities. This list includes almost any possible site of action. For convenience only, these terms are gathered into groups corresponding to body systems. For example, under central nervous system, we have included gross anatomical divisions like cerebrum, hypothalamus and spinal cord, and specific function areas like appetite regulation, emetic center and spinal reflexes. Any document containing information about the emetic or antiemetic action of a drug would be indexed under 'emetic center.' If it also contained information on inhibition of the patellar reflex, it would also be indexed under 'spinal reflexes.' However, simply knowing that there is an effect on the emetic center may not be specific enough. Was the study done in dogs or humans? Was the effect antagonism to apomorphine or motion sickness, or perhaps a study of enzyme systems of the emetic center itself? Was the effect observed only with very high doses? These questions are answered by a second set of terms, or descriptors, any of which can be applied to any word in the main anatomical list. There are words like dog, rodent and human to describe the test subject; words like unusual dose, toxicity and isotope study to describe special circumstances. The really distinctive feature of this system is a group of words describing the actual effect observed, such as action of the drug under consideration (called the *reference drug*) on function, metabolism, or histology of the tissue involved, effect of the reference drug on action or metabolism of some other drug or physical agent.

A combination of these descriptors with the site-of-action words will describe the contents of any pharmacology report. Thus, a paper reporting that very high doses of the *p*-chloro derivative of Boppo antagonize the emetic effect of apomorphine in dogs would be described by the terms: *Site of action:* emetic center; *descriptors:* effect of the reference drug on the action of another drug—dogs—unusual dose—derivative of the reference drug (Boppo).

A paper describing the effect of chlorpromazine on enzyme activity in liver slices would be indexed as: *Sites of action:* liver—tissue enzymes and metabolism; *descriptors:* effect of the reference drug (chlorpromazine) on metabolism—*in vitro* study.

Two years of experience have demonstrated that this subject heading list fills our specifications. It is comprehensive enough to cover any type of drug effect and it is general enough to be applied to any area of pharmacology or physiology. Since each word in it is independent of all the others, it can be expanded and extended at any time.

A carefully planned subject heading list is the most important part of any indexing system. It is the one aspect which absolutely requires the

efforts of a scientist familiar with the field of knowledge involved. Once this list has been prepared, reducing it to punched card codes becomes a mechanical problem. An IBM card has 80 vertical columns numbered from 1 to 80. Each of these columns has 12 punching positions. Every one of the total of 960 positions can be assigned a particular meaning. In normal punched card procedures, only one punching position in any one column should be used at one time. Therefore, each column is usually used to record one piece of information. For example, column 10 could be assigned to species, with position 0 indicating dogs; 1, cats; 2, rats; and so on. A punch in the 11th or so-called x-position of the column might mean that several species are mentioned. Numerical information may be entered as such. For example, columns 18–20 might be used to record dose in mg/kg. In that case, 125 mg/kg would be punched as column 18, position 1, column 19/2, and 20/5. The identification number of the document is often punched in the first or last 6 columns of the card.

Since there are 12 punching positions in each column, our list of less than 260 index headings would require approximately 18 columns of assigned positions, or direct punches. However, very few of these items would be needed to describe any one document, and most of the 18 columns would be blank on every card. Since the same card also had to contain clinical data, document number, author identification, and other general information, all available space had to be used as efficiently as possible. If, in the future, the subject heading list should be expanded, it would require still more space, which would probably not be available. These considerations called for some modification of the conventional methods.

As mentioned above, normal punching procedure does not permit unplanned use of more than one punch in a column. However, it is quite legitimate to plan to use a specified number of punches in each column. If we assign a specific meaning not to each punch in a column but to each combination of punches, the number of possible entries is greatly increased. If, instead of saying that a punch in position 0 indicates dogs, 1 means cats, and so on, we say that the combination of 0 and 1 in the same column is spaniels, 0–2 is beagles, 0–3 is boxers, 1–2 is alley cats, x-9 is human beings, and so on, we now have something like 66 things to choose from instead of 12. If a combination of 6 punches in a column is used, there are 924 possible combinations, that is, 924 possible entries, any one of which can go in one column. If each of the 924 items had been assigned one specific position, we would have had to use 77 of the 80 columns on the card, since each item must always appear in the same place. Using 6-digit codes, we can put any item in any column. We therefore need to reserve only a few columns, enough to contain the maximum number of terms to describe one document.

Our system for storing pharmacological information actually uses both types of punching. We have 4 columns of assigned positions or direct punches for the 48 descriptors, the words that apply to any study such as species, type of action, and special circumstances, since we need to use some of these for every document. We use the 6-digit codes for the site-of-action headings, the anatomical terms. We have found that we seldom need more than 3 or 4 of these terms for any document, but to be safe, we have allowed 7 columns. Remember that any of as many as 924 subjects can go in any of these 7 columns. One additional column of direct punches for body systems has been added, to take care of requests for all documents describing any effect of chlorpromazine on the cardiovascular system, for example. (See APPENDIX for sample portions of the finished code.)

This code sheet containing the subject headings and the codes assigned to them takes care of the input of the system. Entering information in the file requires only two people, a coder and a key-punch operator. Operation of input can be described by following the paper mentioned above, on the antiapomorphine action of *p*-chloro-Boppo. The coder reads the paper and decides which of the code words apply. She circles these words and their accompanying numbers on a code sheet, stamps the sheet 'Boppo,' and passes code sheet and original document on to the key-punch operator. If this is document number 11,539 in the Boppo file, the first 6 columns of the card will be punched 011539. On our cards, columns 7 through 34 are reserved for clinical data, but since the document in question is a pharmacology paper they are left blank. Position 3 in column 35 is punched to indicate that the study involved dogs. Column 36, position 1 indicates that an unusual dose was used; 37/3, that a derivative of Boppo was involved; 38/3, a central nervous system effect, and 39/3, an effect on the action of another drug. Columns 40–

46 are reserved for the site-of-action codes. In this case, the 6-digit code for emetic center (013-689) goes in column 40; the others are left blank. Other punches on the card will indicate that this is a published paper by Jones from the University of Minnesota and that it contains pharmacological data.

The coding of a report on the antihistaminic effect of chlorpromazine in isolated preparations of guinea pig small intestine and tracheal chain will illustrate how a more complex situation is handled. We have made the somewhat arbitrary decision that histamine should be considered a type of hormone and therefore should be included among the main subject headings. Therefore, our 6-digit codes might be column 40/y34689 for histamine, column 41/145678 for small intestine, and column 42/023569 for trachea. The order in which these are entered is unimportant. Any of the three can be in any of those three columns. Since histamine is a main subject heading, column 39 is punched in position 0, effect of chlorpromazine on function, function in this case being the action of histamine. The other descriptor codes are position 7 in column 35 to indicate *in vitro* study, and position 4 in column 38 to indicate the hormones, since histamine is considered to be the main subject of the paper. It should be stressed that the 6-digit codes are not essential to the operation of the system. They are merely a convenient device which allows us to tie up only a small part of the card, while maintaining a very large reservoir of index entries from which to choose.

These examples illustrate the use of the code to prepare our reference file. The machine we use to search this file is IBM's Model 101 Electronic Statistical Machine. This machine was designed originally to handle census information, and is a big brother to the sorter that picks out the questions on cooking or Abraham Lincoln on the television quiz program. It is not a computer or a giant brain. All it can do is sort and count. Unlike its little brother, it can also do a limited amount of printing and, most important, it can sort for several criteria at a time. The standard sorter can read and sort on only one column, that is, for one characteristic at a time. If you wish to look for a second factor, the cards must be run through a second time. The 101 reads all the columns on a card in one pass and can therefore be instructed to pick out cards which contain several criteria at once. Instructions for each search are given to the machine through a control panel wired by the operator. The wiring procedure is relatively simple, and need not be described here. The important thing is that the machine will separate from a large deck of cards those which contain any desired combination of characteristics. These characteristics include both the presence and the absence of certain punches; that is, we can sort out all cards which combine characteristics A, B, and C, or all cards which contain characteristics A and B only if C is not present. For example, we can obtain all cards for documents which report antagonism to the emetic action of drugs in dogs. We can also specify that they should not contain any reference to sedative effect.

These three factors—the subject heading list, the coding procedure and the machine—compose our system. When a request comes from one of our scientists for all the information on some aspect of chlorpromazine pharmacology, the supervisor checks the code sheet for the appropriate codes. The machine operator wires the control panel according to these codes, and runs the cards through the machine. Since the document number is typed on each of the cards, those which have been selected can be examined visually, or the machine can be instructed to produce a printed list of document numbers. The corresponding documents are taken from the file for detailed study by the original requestor, or for abstracting by someone in the Science Information Department. If the request was for a bibliography, perhaps to be sent to someone outside the company, this can be prepared automatically. This is done on the Flexowriter, an automatic electric typewriter controlled by perforated paper tapes. One of these tapes has been prepared for each reference in the chlorpromazine bibliography. To prepare the special bibliography, the tapes for the references selected on the 101 are fed into the Flexowriter which types them out rapidly and accurately. Once the tapes have been prepared, the references can be reproduced as many times as desired, without the necessity for further proofreading.

The details of the system as we have described them here were designed to fit our equipment and our needs, and are only a small part of a larger operation. The method could, however, be readily adapted to a much smaller operation and much less expensive equipment. The same type of code sheet could be readily adapted to edge-notched cards, such as McBee cards, to eliminate the machines altogether. The most

important part of the whole system, from the pharmacologist's point of view, is the basic philosophy used in designing the subject heading list; that is, the idea of a set of organ and tissue words combined with a second set of words that indicate whether the effect is on function, metabolism or structure of the organ or on the action of another agent on it.

There are several advantages of our system over other methods of indexing:

*1*) It requires no decisions by the indexer about the relative importance of various factors. Considering a paper on epinephrine reversal by chlorpromazine, no one has to decide whether the significant entry should be epinephrine, chlorpromazine, pressor response, hypertension, or adrenolytic action. All these facets of information are coded. No matter which is used as a search criterion, the paper will be found.

*2*) Once the subject heading list has been set up, highly trained scientists are not needed to operate the system. The indexers must have scientific training, of course, since they must be able to read reports intelligently and translate the authors' words into the code words we use. An occasional arbitrary decision as to where something should be coded has been necessary. These arbitrary decisions do not decrease the accuracy of coding or retrieval, since everyone who uses the system knows which concepts are included under each code word.

*3*) Uniformity of indexing is very high. We have tested this several times by having different people index the same papers and the results have always been very gratifying.

*4*) Relative speed of searching is inherent in machine systems. No time is lost deciding where to look for something or in hunting down 'see also' references. The IBM 101 examines cards at a rate of 450 per minute, or 9,000 in 20 minutes. Refiling of cards is unnecessary. They can be replaced in the file drawer in any order without regard to alphabetical or numerical arrangement. The only requirement is that they all be the same side up!

*5*) The subject heading list can be enlarged or made more detailed any time we think it is necessary without disturbing indexing already done, and without upsetting the entire classification system. If necessary, we can sort out cards referring to any particular subject and add more codes to them. The punched cards are very easily reproduced mechanically if extra decks are desired.

*6*) The code was planned to be applicable to any area of pharmacology or physiology. As a secondary benefit, we have found it useful for such things as indicating fields of interests of consultants, and even for describing side effects in patients.

This is, of course, a retrieval system. It does not give you the information you want, but it tells you which documents in the file contain it. We have found that this is the most satisfactory way to deal with pharmacological information. For handling clinical information, the code is set up with a high proportion of direct punches or assigned positions, so that we can also correlate and tabulate data. This is done by wiring the control panel to instruct the 101 to count the cards as it sorts them. The tables are then printed out by the machine just as we print out the document numbers.

It is important to realize that this type of index is not limited to the equipment which we use. It was designed for use with the IBM 101 sorter, but changing the code numbers, not the words, could readily adapt it to a simple sorter or to hand-sorted cards. It could probably also be adapted to a card index. Needless to say, it would also be applicable to other types of machine systems. The same philosophy could also be used in designing similar indexes for other fields, such as microbiology.

This unique system of indexing and searching pharmacological literature provides the scientist with several benefits. It frees him from many hours of tedious library searching. It also assures him of uniform, comprehensive analysis of all documents in his sphere of interest. At his request, he can obtain all the documents, and only the documents, which contain the information he desires. Such service, we feel, is an important factor in the most efficient and productive use of scientific manpower.

## APPENDIX

### Selected Portions of Pharmacology Code Sheet

| Descriptors | Sites of Action |
|---|---|
| *Column 35:* Subject | *Central nervous system* |
| 0 not specified, several | 013456 cerebrum |
| 1 rodents | 013457 hypothalamus |
| 2 birds, amphibia | 013458 brain stem |
| 3 dog and cat | 013459 spinal cord |
| 4 monkey | 013578 temperature reg. |
| 5 other animal | 013579 appetite centers |
| 6 human | 013678 respiratory centers |
| 7 *in vitro* | 013679 vasomotor centers |
| 8 ovum | 013689 emetic centers |
| 9 pregnant | 014567 spinal reflexes |
| x fetus, newborn | *Gastrointestinal system* |
| y other | 134789 stomach |
| *Column 36:* Special features | 135678 gastric motility |
| 0 unusual dose | 135679 acid secretion |
| 1 unusual route | 135689 other secretion |
| 9 absorption | 135789 small intestine |
| x combinations | 145678 s. i. motility |
| y comparisons | 145679 s. i. absorption |
| *Column 37:* Type of study | 145689 s. i. secretion |
| 0 chemical | 145789 colon |
| 1 bioassay | 146789 color absorption |
| 2 pharmacy | 156789 colon motility |
| 3 derivative of RD* | *Hormonal actions* |
| 4 isotope study RD | y12345 ACTH |
| 5 isotope study OD† | y12346 TSH |
| y multiple action | y12347 STH |
| *Column 38:* Body system | y12348 other anterior pituitary |
| 0 metabolism of RD | y12459 estrogens |
| 1 blood, hemopoietic | y12467 progesterone |
| 2 cardiovascular | y12468 other female |
| 3 central nervous | y12469 androgens |
| 4 endocrine, hormone | y12567 other male |
| 5 gastrointestinal | y23456 adrenocortical |
| 6 peripheral nervous | y23457 glucocorticoids |
| 7 urogenital | y23458 cortisone |
| 8 respiratory | y23459 hydrocortisone |
| 9 bone, muscle, skin | y34678 adrenal medullary |
| x body in general | y34689 histamine |
| *Column 39:* Type of action | y34789 ACh |
| 0 RD on function | |
| 1 RD on metabolism | |
| 2 RD on histology | |
| 3 RD on action OD | |
| 4 RD on metabolism OD | |
| 5 RD on action PA‡ | |
| 6 OD on action RD | |
| 7 OD on metabolism RD | |
| 8 PA on action RD | |
| 9 comparison with PA | |

* RD—Reference drug.

† OD—Other drug. ‡ PA—Physical agent.