ARTICLES

The NCI Drug Information System. 1. System Overview

G. W. A. MILNE*

Information Technology Branch, Division of Cancer Treatment, National Cancer Institute, Bethesda, Maryland 20205

J. A. MILLER

Fein-Marquart Associates, Baltimore, Maryland 21212

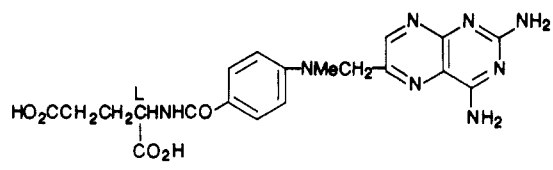
Received April 21, 1986

An interactive computer system has been designed to handle all the data associated with the National Cancer Institute's (NCI) drug screening program. The system resides on the NIH DEC System 10 computers and allows interactive access to the entire NCI screening data system. This contains over 20 separate databases, including a chemistry file of about 400 000 structures and a biology file of approximately 1.5 million test records. New compounds and test data are added daily to the files, and the system also controls and records all the daily operations of the screening program, such as acquisition, shipping, and biological testing of chemicals.

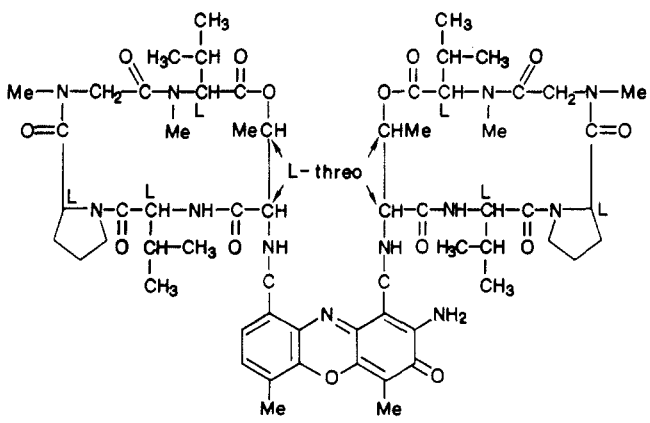
INTRODUCTION

In 1955, the National Cancer Institute (NCI), a part of the National Institutes of Health, embarked upon a program in which chemical compounds are examined for activity against various forms of cancer.[1] In the years prior to 1955, a number of compounds had been found to possess therapeutic activity against cancer. These included mustard gases such as nitrogen mustard (I),[2] antifolic acid compounds such as methotrexate (II),[3] and antibiotics such as actinomycin D (III).[4] These

I



II



III

compounds are usually somewhat toxic toward normal cells, but more toxic toward cancer cells, and by means of this differential cytotoxicity, they offer a therapeutic advantage to the cancer patient.

In 1955, it was known only that various structurally diverse chemicals, such as those cited above, exhibited such anticancer activity. The subtle relationships between chemical structure and differential cytotoxicity were no more understood then than they are today, however, and for this reason, it was decided that broad, generally unbiased, screening for anticancer activity of large numbers of chemicals would be the method of choice. As experience has accrued, some relationships between structure and activity have been discerned, and information of this sort is now being used in attempts to select more promising candidates for the program.

Some of the policies adopted in 1955 are particularly important because they are still in effect and they impact directly upon contemporary information policy in the NCI. First, the number of chemicals acquired and tested annually was approximately 10 000. In the 30 years of the program, this number has fluctuated, but the 30-year average is 13 000 new compounds per year. Currently, the number of new compounds tested per year is again set at 10 000. Such high numbers of test compounds are a direct result of the low probabilities of discovering useful antitumor activity.[5] Second, from the outset, the tested compounds were provided to NCI as free gifts. This was a largely practical decision, which led, however, to a policy under which NCI would, if requested, accept such gifts under conditions of confidentiality. This was, and still is, necessary in order that compounds may be submitted to the program by for-profit organizations. Such sources include pharmaceutical and chemical companies, whose contribution could be significant, but who require confidentiality in order to retain ownership rights to chemicals not yet protected by patenting. Third, in 1963, it became clear the system must track not just individual chemical compounds, but distinct samples of chemical compounds. There is typically more than one sample for any chemical, and consequently, this new policy had the effect of magnifying the data management problem considerably.

By these three decisions some major constraints were established and continue today to control some part of the operation of data handling at NCI. As a result of the first decision, the database is large and growing quite steadily. The situation that results from the second policy is that approximately half the database is confidential. The third point has the effect of creating a new database, the Inventory database, which has more records than the Chemistry database. Each of these factors has had to be accommodated in the implementation of data handling procedures.

Most of the record keeping in the first decade of the program was manual. Searching of files was difficult or impossible, and much reliance was placed upon institutional memory and paper records. As the databases grew larger, such expedients became untenable and in the mid-1960s, both the chemistry files and the biology files were entered into different computer systems. The biology files resided in an IBM

0095-2338/86/1626-0154$01.50/0

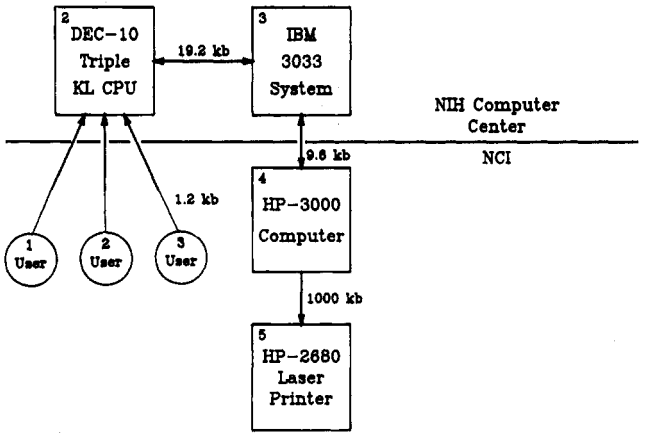


Figure 1. Computers used by the DIS.

360/370 batch computing environment at NIH until recently. The chemistry files were loaded into computers at Chemical Abstracts Service.[6] There, over a 15-year period, an interactive search capability was gradually developed and proved, in fact, to be one of the forerunners to the CAS ONLINE system.[7]

The dichotomous nature of this arrangement proved to be increasingly difficult and inefficient. In 1979, NCI convened a group of experts to review the situation and received in return the recommendation[8] that these large databases be consolidated in an interactive environment, preferably on an NIH-controlled computer because of the data security requirements. The advice was adopted, and in late 1981, a plan was framed to move in this direction. This plan was embarked upon in 1982 with the award of a contract to Fein-Marquart Associates for the design and development of an integrated system to support these activities. The system that emerged in stages during late 1984 and early 1985 has been named the NCI Drug Information System (DIS) and is described in detail in this and the companion papers in this series.

DESIGN REQUIREMENTS

The basic task of the NCI drug screening program is to acquire small quantities of many chemical compounds and test them for activity against cancer. The DIS provides the entire data processing support for this program, and the backdrop against which the DIS works is described in this section.

1. The NIH Computer Center. In accordance with the recommendation of the reviewing committee, the DIS was to be developed and installed on computers of the NIH Computer Center. This is a fairly large computing environment that has been described elsewhere,[9] but a very brief description of the center will be provided here by way of context. Two mainframe systems are provided at the NIH Computer Center, and their relationship is shown in Figure 1. These are a large IBM System 3033 (3 in Figure 1) and a smaller DEC System 10 (2 in Figure 1). Both systems are accessed primarily from remote dial-up terminals (1 in Figure 1). The two computers are themselves connected so that data can be passed at high speed from one machine to the other under a user's control. Virtually all interactive access to the DIS by users (1 in Figure 1) involves the DEC-10. The major role of the IBM mainframe is to provide support in data processing, particularly in database updating and output printing. The DIS also makes heavy use of a Hewlett-Packard 2380 laser printer (5 in Figure 1), which is controlled by an HP-3000 processor (4 in Figure 1), the pair being configured as an output peripheral to the IBM computer. All data destined for this printer, both graphics and text, is passed from the DEC-10 via the IBM system. The computers support many languages, but the DIS is written almost entirely in FORTRAN. About 5% of the DIS code is written in machine-language. This is a fairly complex, but very powerful environment because the two mainframes offer different strengths, and by sending work to one computer or the other, the DIS can take advantage of these differing strengths as well as the redundancy that is implicit in the arrangement.

2. Management Capabilities. The entire drug screening effort at the NCI has been designed as a production line in which the various steps through which a compound passes are sequential. Acquisition, of necessity, must precede the establishment of a material inventory, and this is done before testing begins. Each of these tasks is handled by contractors to the government, and it is therefore important that each of them proceeds at an optimal rate or else subsequent steps will be hampered and impact on later contracts could be significant. A primary requirement of the DIS, therefore, is that it allow timing and control by NCI staff of all the individual operations which together comprise the drug screening program.

Such control must be available upon an interactive basis in real time. A computer system which permitted only ex post facto access to the data would be inadequate in that it would not allow any forward planning. Two implications that flow from the requirement for interactive control are that a generalized interactive search capability be provided and that the databases be kept as current as possible. These are both major design considerations and will be expanded upon in some detail below.

3. Database Searching Capabilities. The information necessary for program management tends to be detailed and practical. Managers need to know, for example, the current load on a particular screening laboratory so they can determine if more compounds should be shipped there or elsewhere. Similarly, when secondary or follow-up testing of a compound begins, the details of the primary screening, which was carried out at a different laboratory, must be readily available. The DIS must therefore allow rapid access to data of this sort, and wherever it is feasible, the system must be able to make routine operating decisions on an automatic basis.

At the same time, however, there is a significant need for a broad and comprehensive searching capability which will permit senior staff to attain database-wide perspectives concerning questions such as the types of chemical structures that have been tested or the long-term performance associated with different approaches to problems.

4. Database Updating. The DIS uses an inverted file structure in order to provide an interactive search capability. This deals satisfactorily with the search requirements, as has been seen elsewhere,[10] but complicates updating because, in principal, when data are added to the database, a complete reinversion is necessary. With databases as large as those in the DIS, such reinversions are prohibitively expensive, and so a different approach is necessary. A solution to this problem is to invert database increments independently of the main database. The pointer files must then be adjusted to reference not only the main database, but all the incremental additions. As the number of increments increases, search times will also increase, and at some point it becomes more cost effective to consolidate all the increments into one large increment or into the main database.

New data are entered into the DIS continually by as many as 40 people associated with the various program functions such as acquisition, inventory, shipping, and screening. Three of these people, as is discussed in the next paper in this series, work full-time entering chemical structures into the DIS, but for all other people DIS data entry is only a part of their overall responsibilities and often occupies less than 10% of their time. The data may be displayed as soon as they have been entered, but they only become searchable after they have been processed through an update. Incremental updating allows DIS

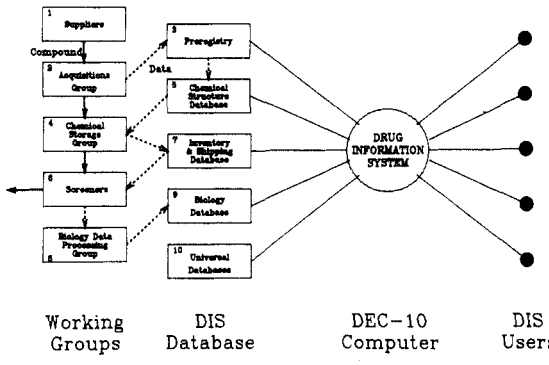


Figure 2. Operating structure of the DIS.

database updates to be performed as frequently as daily, although some of the databases are updated only once per week. Quite large numbers of increments can be tolerated. In the Chemistry database, for example, 30 increments representing 60 weeks' worth of new data, or some 12 000 new structures, have been added with no serious degradation in search times.

Updating is generally a batch process, controlled by programs that start automatically at predetermined times and that, once the update has been completed, resubmit themselves for the following day, week, or fortnight, depending on the database. Updating of the different DIS databases is generally carried out between midnight and 6:00 a.m. when no users are present. As an added precaution, the updating programs prevent user access to the database during its update processing.

OVERALL DESIGN

The primary requirements for the DIS were defined in the previous section. The system must run within the NIH Computer Center, and it must provide interactive and rapid access to data at a high level of currency. These constraints led to the design that is described in this section.

1. DIS Operation. The operation of the drug screening program and the DIS is illustrated in Figure 2. The acquisition step (2 in Figure 2) represents the first DIS operation for a compound. From a variety of sources, including literature surveillance and liaison with industry and academia, the program identifies each year some 50 000 structures judged to be potentially of interest in connection with cancer chemotherapy.

These structures are all provided to the Acquisitions group, which enters them into the Pre-Registry database of the DIS. Using criteria such as structural uniqueness and estimated probability of activity,[11] computer programs identify about 20 000 of the better candidates from this input, and the Acquisitions group requests samples of these compounds for testing.

Approximately half the compounds requested are received. As they are acquired, they are assigned a permanent "NSC Number",[12] and their chemistry records are moved from the Pre-Registry to the Chemistry database. The physical samples are labeled with their NSC Numbers, in barcoded form, and transferred to a storage facility, where they are logged in. This second contractor weighs the material and creates an Inventory record for that sample. A Shipping record is also begun at this point, reflecting the fact that the compound was shipped on given dates from the Supplier to the Acquisitions contractor and from the Acquisitions to the Storage contractor. These new records are used respectively to update the Inventory database and the Shipping History database.

For preliminary testing, which is against P388 leukemia in mice, the DIS controls the flow of compounds from the Storage contractor to the various Screening contractors.[13] As a screener's load/capacity ratio drops, the DIS automatically directs more compounds to be sent to that screener. The capacity of a screener can be adjusted by NCI staff to reflect the screener's contractual obligation. The storage contractor receives such shipping requests from the DIS and fills them on a daily basis. Each year, there are some 10 000 such "automatic shipments", and in addition, some 2000 individual shipments of compounds destined for secondary testing are ordered by NCI staff.

Table I

database	no. of records (approx)	no. of fields
Chemistry	410 000	52
Inventory	498 000	44
Shipping	315 000	18
Biology	1 300 000	101
Namecodes	8 000	18
Pre-Registry	20 000	87
Order	11 000	26
Test Vehicles	50	3
Acquisition Categories	25	2
Geographic Categories	273	4
Test Systems	524	5

The Screening laboratories use a full-screen edit program operating on a Hewlett-Packard HP-2645A terminal to collect the data from completed screening experiments (6 in Figure 2). Once all the data have been entered, they are written in condensed form onto a tape cassette in the terminal. At regular intervals, typically daily, the terminal is logged onto the NIH computer facility, and the contents of the tape are downloaded into the NIH IBM 370 computers. There, the downloaded files are used as input to a program (8 in Figure 2) that examines all new data for internal consistency and freedom from logical errors and then calculates the final test results from the raw data. Errors that can be corrected on the spot are resolved; other errors that are detected are passed back to the screener. When the screener logs on next, the calculated data and the errors are presented for resolution before more data entry begins. When data have been finally validated in this way, they are written to a staging area to await the next master file update. These updates are carried out every two weeks and trigger an update of the online searchable files in the DIS. Such a biweekly update is reflected in the content of the searchable biology database shown as 9 in Figure 2.

2. DIS Database Elements. Each of the steps described in Figure 2 results in a modification of some part of the overall database of the program. Since this database is so large, it is divided into some 24 distinct subsidiary databases. Each user of the system has controlled read/write privileges: typically, a member of the Acquisitions group can read the Inventory database but cannot write in it. That privilege is reserved for members of the Storage group, who in turn cannot write in the Pre-Registry database.

Of the 24 databases, 11 are searchable. These are listed in Table I. Each of the databases contains some number of fields, and each field is identified by means of a "field mnemonic", which is usually a four-letter code such as ADDR for address or MOLF for molecular formula. There are 360 distinct fields in the DIS; 232 of these are searchable, and all of them can be displayed on command. Every one of these field mnemonics is unique; it is therefore unnecessary for a user to remember which database is being addressed, because the DIS can recognize the field mnemonic and search the appropriate database.

CAPABILITIES

In this section, a brief review of the flow of data in, and the search capabilities of, the DIS is provided. No detail will be

given; for such information, the reader is referred to the appropriate paper of this series.

1. Data Input. When a structure is considered for acquisition and testing, it is first entered into the DIS. Several specialized programs have been written for this task, and those currently in use run on microcomputers which also can behave as terminals to the mainframe. A graphics representation of the structure is entered into the microcomputer, operating in a stand-alone mode. When the entry is complete, this graphical representation is reduced to a set of vectors. After a series of data evaluation and validation checks have been successfully performed in the microcomputer, the vector set is converted to a connection table and both the set and the table are transmitted to a staging file in the mainframe ready for updating into a master file.

The connectivity table, which, aside from stereochemistry, is a full description of the structure, is used as the source of all the submolecular fragments used for structure searching. The vector set, which contains stereochemical information, is retained with the express purpose of providing a high-quality structure diagram at output time.

The level of activity of the drug screening program requires that some 30 000 structures be entered each year. This has led in turn to considerable study as to the best structure input method because less than 5 min can be budgeted to entering and checking each structure. The various methods examined and the method selected are described in more detail in a subsequent paper in this series.

2. Search Capabilities. All data in the DIS fall into one of two categories. The bulk of the Chemistry database—about 25% of all the DIS data—consists of information pertaining to the chemical structures of the approximately 400 000 chemicals that have entered the program since 1955. The remaining 75% is entirely alphanumeric, with items ranging from those consisting of a single character to those represented by lengthy character strings. Searching through the alphanumeric data is managed by means of a database management system designed for this purpose and known as TDRS.[14] Structure and substructure searching within the Chemistry database is carried out by using an enhanced version of the software from the NIH–EPA Chemical Information System.[15] All DIS searches lead to the creation of results files, and all results files, independent of their origin, are compatible with one another and can be merged and intersected.

The DIS databases are managed by a software package known as TDRS (Text/Data Retrieval System). This package provides a multitude of capabilities for dealing with databases comprising a series of data fields containing textual, numeric, and data information. Within the framework of TDRS, provisions are also made for dealing with data fields containing other special information, such as the chemical structure information so important to the DIS. To perform a general search in the DIS, the field to be searched must be identified, but the user is not required to designate the database. All DIS fields have unique mnemonics, and the user may ignore the fact, which is largely transparent, that the DIS contains many databases. A search expression may be very simple, taking the form

```
OPTION? ADDR/CALIFORNIA
```

or it may contain ranges or inequalities

```
OPTION? DACQ/5-JAN-81 TO 14-APR-81
OPTION? NSC/>382500
```

and Boolean operators are permitted within the search expression

```
OPTION? (DACQ/>1-JAN-85 OR NSC/>600000) AND
CAMT/>500 MG
```

This capability allows the user either to retrieve precisely the information required or to query the database iteratively, narrowing down the number of retrievals with successively more specific search statements.

The Chemistry database contains a considerable amount of data that can be searched in the standard manner by TDRS. Thus the molecular weight (MW), chemical name (CNAM), NSC Number (NSC), and amount of compound (AMT), among other fields, may be searched by using the field/value format. In dealing with the chemical structure data, however, the field/data format cannot be used because structures and substructures are not readily represented by character strings and the DIS defers to the structural search package. Attempting a search of this sort, the user is required to define a structural fragment of interest and then to query the database in terms of this fragment. Structure definition is accomplished by means of a set of structure building commands.[15] These allow the generation of rings or chains of atoms, addition of branches, and specification of atom and bond types.

Given a query structure, the user can retrieve from the database any compounds whose structures exactly match that of the query structure. Alternatively, searches may be done for compounds containing a specific "atom-centered fragment" (an atom with one or more neighbors specified) or a specific ring or ring system. As in TDRS searching, all these structure searches lead to results files that can be further refined by merging or intersection with each other. Thus within the framework of the DIS, it is possible to evaluate very broad queries such as "all compounds with one or more 6-membered rings" or very precise requests such as "all compounds supplied by a given corporation on a given date and containing a specific substructure but with no more than 10 carbon atoms". Both the standard TDRS searching and the special chemical structural searching are complex subjects, and they are dealt with in the necessary detail in the subsequent papers of this series.

3. Data Output. For a system as large and complex as the DIS, control of data output is a major task. The type and amount of data that are implied in an output command are controlled entirely by the user, and data may come from any of the databases. Data moreover may be required to be presented in a user-specified order and format, and the identity, location, and nature of the output device may also be specified by the user. A further complication is that if the output is, as is very common, to contain one or more chemical structures, then a standard high-speed printer is no longer the optimal output device. A final difficulty is that a major type of DIS output consists of letters to suppliers of chemical compounds. These require NIH letterhead and a signature and may have to be written in a language other than English. They need not therefore use the Roman alphabet: minor deviations, to French and German, are common, as is the fundamental switch to Japanese.

For reasons of this sort, the DIS makes heavy use of a high-speed laser printer with full graphics capability (5 in Figure 1). This machine can handle any of the data types that the DIS requires and represents the method of choice for much DIS output. Users of the system are not, however, required to use the laser printer. Output can be directed to a nongraphics terminal, a graphics terminal, or a nongraphics high-speed printer such as the IBM 3800. As one moves further afield from true graphics devices, the quality of the graphics output suffers, and while chemical structures can be represented on nongraphics printers, there are, as can be seen from Figure 3, some major compromises necessary.

4. Data Security. Fully half the compounds in the DIS have been donated to the government under an agreement that the government will keep the data concerning the compounds

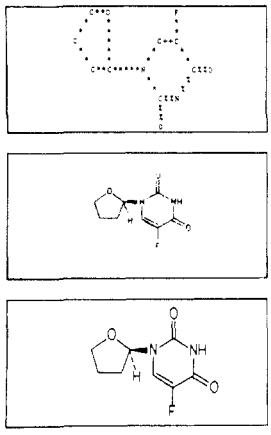


Figure 3. Different types of structure output from the DIS.

confidential. The confidence that such suppliers place in the NCI is very important; consequently, a major effort has been made in the DIS to respect and maintain this confidentiality. A second type of security that is important in the DIS concerns the ability of a user to write in DIS files. A single security system has been built into the DIS to cope with both these problems, and this is described briefly here.

Every compound and every sample of every compound in the DIS is classified as either "open" or "confidential".[16] These classification flags themselves constitute one of the DIS databases; it is a database that is closed to all but a very small number of people and which is reviewed and updated on a weekly basis.

Any user of the NIH computer may attempt to access the DIS, but access is permitted only to those whose identification is explicitly known to the DIS. Within the group of DIS users, many levels of privilege are possible. One user may be authorized to view any DIS data, but another may only be allowed to see nonconfidential data. Further, while two users may both be allowed to view confidential chemical structures, it may well be that only one of these individuals is allowed to alter a chemical structure in the database.

When any person accesses the DIS, a DIS controller program notes their identity and reviews their privilege level. Every command subsequently issued by that person is weighed by the controller in the light of this privilege level. Only when it is clear that the user's authority will allow whatever activity is implicit in the command is the command honored. If the command is a search which retrieves confidential data and the user is not authorized to view such data, the DIS detects this and removes all confidential data from the search results file before informing the user as to the completion of, and results from, the search. Every response the DIS makes to a user is first examined in terms of the user's authority. In this way, it is very easy to detect and prevent direct attempts to penetrate the database, and tests have shown that extremely indirect approaches, such as teasing out confidential data by means of complex Boolean strategies, are likewise precluded.

SUMMARY

An intensive effort over a period of 4 years has resulted in the design, development, and implementation in the National Cancer Institute of a Drug Information System.

This system runs on a DEC System 10 computer and supports the entire data management aspect of NCI's drug development program. All modules of the DIS are used interactively, and control over all aspects of the program can be maintained through the DIS.

The databases which comprise the DIS include a file of over 400 000 chemical structures, together with their biology test records and a large amount of management data that has been accumulated during the 30 years this program has been in existence. The entire storage requirement of the DIS is slightly less than 4 Gbyte.

Development of the DIS took between 3 and 4 years and required about 25 man-years, most of which was accounted for by programmers and systems analysts. The total contractual cost of this work was approximately $1.3 million, and the computing cost was close to $2 million.

Continuing costs of the system are greatly reduced. Database maintenance, which is a very active area, costs perhaps $200 000 per year, and currently a like amount is budgeted for software enhancements, many of which are not absolute requirements but which represent efforts to improve the user interface with the system.

REFERENCES AND NOTES

(1) The starting point for this effort is generally regarded as the Congressional authorization, in 1955, of $5 million for the establishment by NCI of a Drug Development Program. See: DeVita, V. T.; Oliverio, V. T.; Muggia, F. M.; Wiernik, P. W.; Ziegler, J.; Goldin, A.; Rubin, D.; Henney, J.; Schepartz, S. "The Drug Development and Clinical Trials Programs of the Division of Cancer Treatment, National Cancer Institute". *Cancer Clin. Trials* **1979**, *2*, 195–216.

(2) Gilman, A.; Philips, F. S. "The Biological Actions and Therapeutic Applications of β-Chloroethylamines and Sulfides". *Science, (Washington, D.C.)* **1946**, *103*, 409–415. See also: Ross, W. C. J. "Rational Design of Alkylating Agents". In *Antineoplastic and Immunosuppressive Agents. I*; Sartorelli, A. C., Johns, D. G., Eds.; Springer Verlag: New York, 1974.

(3) Seeger, D. R.; Cosulich, D. B.; Smith, J. M.; Hultquist, M. E. "Analogs of Pteroylglutamic Acid. III. 4-Amino Derivatives". *J. Am. Chem. Soc.* **1949**, *71*, 1753–1758. See also: Mead, J. A. R. "Rational Design of Folic Acid Antagonists". In *Antineoplastic and Immunosuppressive Agents, I*; Sartorelli, A. C., Johns, D. G., Eds.; Springer Verlag: New York, 1974.

(4) Waksman, S.; Woodruff, H. B. "Bacteriostatic and Bacteriocidal Substances Produced by a Soil Actinomyces". *Proc. Soc. Exp. Biol. Med.* **1940**, *45*, 609–614.

(5) Over the lifetime of the program, 1 compound from every 3000–5000 tested has shown sufficient activity to enter clinical trials, and about one-tenth as many have reached the point of being marketed as drugs for the treatment of cancer. There are currently 39 commercially available drugs used in cancer treatment, but 9 of these are considered to be hormones. Twelve of the thirty cytotoxic drugs were developed by NCI, which was also involved at some stage in the development of most of the remaining eighteen.

(6) Richman, S.; Hazard, G. F.; Kalikow, A. K. "The Drug Research and Development Chemical Information System of NCI's Developmental Therapeutics Program". In *Retrieval of Medicinal Chemical Information*; Howe, W. J., Milne, M. M., Pennell, A. F., Eds.; ACS Symposium Series 84; American Chemical Society: Washington, DC, 1978; pp 200–221.

(7) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. "The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.

(8) Report of the NCI Ad-Hoc Extramural Committee on Information Management, Jacobus, D. P., Chairman, Feb 1979.

(9) The overall design of the NIH Computer Center changes continually. A recent description of the various computer systems in this center will be found in the NIH-DCRT publication, "Interface", No. 127 (25 Dec, 1985). Copies of this publication are available from the Division of Computer Research & Technology, NIH.

(10) Heller, S. R. "Conversational Mass Spectral Retrieval System and Its Use as an Aid in Structure Determination". *Anal. Chem.* **1972**, *44*, 1951–1961. Feldmann, R.; Milne, G. W. A.; Heller, S. R.; Fein, A.; Miller, J. A.; Koch, B. "An Interactive Substructure Search System". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 157–163.

(11) The programs that are used to provide estimates of probable antileukemia activity are those developed by Hodes. See: Hodes, L. J. "Computer-Aided Selection of Compounds for Antitumor Screening: Validation of a Statistical-Heuristic Method". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 128–132. Hodes, L. J. "Selection of Molecular Fragment Features for Structure–Activity Studies in Antitumor Screening". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 132–136. See also: Hodes, L. J.; Hazard, G. F.; Geran, R. I.; Richman, S. "A Statistical-Heuristic Method for Automated Selection of Drugs for Screening". *J. Med. Chem.* **1977**, *20*, 469–475. Hodes, L. J. "Computer-Aided Selection of Novel Antitumor Drugs for Animal Screening". *ACS Symp. Ser.* **1979**, *112*, 583–603.

(12) The acronym "NSC" stands for National Service Center, a short form of Cancer Chemotherapy National Service Center, the early name for the program. The "NSC Number" is used by NCI as a Registry Number. Other Registry Numbers, such as the CAS Registry Number

are not useful for NCI because CAS Registry Numbers cannot be assigned to the confidential structures which constitute about half of the NCI database.

(13) The Screening Laboratories that are currently under contract to NCI are The Southern Research Institute (Birmingham, AL), Battelle Columbus Labs (Columbus, OH), Illinois Institute of Technology Research Institute (Chicago, IL), Mason Research Institute (Worcester, MA), Institut Jules Bordet (Brussels, Belgium), Arizona State University (Tucson, AZ), and the University of California (Los Angeles, CA).

(14) Heller, S. R.; Milne, G. W. A.; Feldmann, R. J. "A Computer-Based Chemical Information System". *Science (Washington, D.C.)* **1977**, *195*, 253–259.

(15) Milne, G. W. A.; Heller, S. R.; Fein, A. E.; Frees, E.; Marquart, R.; McGill, J. A.; Miller, J. A.; Spiers, D. S. "The NIH–EPA Structure and Nomenclature Search System". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 181–186.

(16) Confidential compounds in the DIS are described as "discreet", and their NSC Numbers are prefixed with a "D". The misuse of this word has long been recognized; it is a practice, however, that is so ingrained that no amount of reference to dictionaries can be expected to change it.

The NCI Drug Information System. 2. DIS Pre-Registry

G. W. A MILNE* and ALFRED FELDMAN

Information Technology Branch, Division of Cancer Treatment, National Cancer Institute, Bethesda, Maryland 20205

J. A. MILLER, G. P. DALY, and M. J. HAMMEL

Fein-Marquart Associates, Baltimore, Maryland 21212

Received April 21, 1986

The Pre-Registry Module of the Drug Information System (DIS) is a staging area through which all new compounds are passed prior to acquisition and testing. Several methods are available for the entry of structures into the Pre-Registry; all involve built-in data validation. Newly entered structures are examined by computer programs for structural novelty and potential for anticancer activity. For those compounds that proceed to acquisition, the various acquisition steps, such as letter writing and record updating, are performed automatically. When a sample is obtained, the entire Pre-Registry record is updated and moved forward into the permanent DIS chemistry files.

INTRODUCTION

The goal of the Developmental Therapeutics Program at the National Cancer Institute (NCI) is to identify compounds that possess utility in the treatment of human cancer. In order to identify promising compounds, the program maintains liaisons with several thousand research organizations around the world and also supports two literature surveillance efforts, one focused on natural products and the other on synthetic chemicals. All the data associated with these activities are stored in the NCI Drug Information System (DIS), which is described in this series of papers.

When a promising compound comes to the attention of the NCI drug screening program, its structure and other relevant information are entered into the Pre-Registry subsystem of the DIS. Upon further examination, only about 40% of these structures are determined to be of sufficient interest to merit testing, and with only a fraction of these, i.e., some 30% of the original entries, are acquisition efforts successful. The Pre-Registry is therefore a staging file, and compounds in the Pre-Registry are only assigned temporary identification (TID) numbers. No more than 30% of all Pre-Registry entries are subsequently passed through to the permanent files of the DIS where each is assigned a permanent identification number.

As a first step then, the Pre-Registry must ascertain that a compound is new to the program. Compounds that have been tested are not reacquired.[1] If the compound is new, it must further be determined whether there is any reasonable expectation that it may have any activity against cancer. If the compound is not eliminated by either of these tests, a decision is made as to whether or not it should be acquired. If this decision is affirmative, the Pre-Registry sets the acquisition process in motion.

When the acquisition is complete and a sample of the compound has been received, the programs must log in the sample and acknowledge its receipt. At this point, the Pre-Registry record is moved to the permanent DIS files and the compound is assigned an NSC Number.[2] The actual sample, meanwhile, is forwarded to a storage facility[3] which is responsible for inventory and shipping data, as is described in part 4 of this series. A housekeeping task for the Pre-Registry involves the disposition of all the records for old compounds, whether they be selected and acquired, selected and not acquired, or not selected.

SOURCE OF CHEMICALS

The screening of chemicals for antitumor activity has been continuing at NCI since 1955. In the early years of the program, there were many compounds available and of interest in connection with this screening effort. As a result, selection and acquisition of compounds was not a major task. The screening capacity of the program, however, has always been considerable, and to date, over 400 000 distinct compounds have been tested.

A substantial proportion of the compounds tested by NCI have never been published and therefore are not included in the 7.9 million compounds registered[4] by Chemical Abstracts Service (CAS). It follows therefore that NCI has examined less than 10% of all published structures, and there are some 7 million published compounds which have not been tested. In spite of this, it is no trivial task for the NCI to find thousands of compounds that are (a) available in half-gram quantities and (b) not closely related chemically to compounds which have already been tested.

The funding and staffing of the NCI program currently allow the screening of about 10 000 compounds per year. It is a matter of experience that of every three to five compounds considered for testing, only one will actually be acquired, and accordingly the program considers at least 30 000 compounds

0095-2338/86/1626-0159$01.50/0