turers were involved in developing the product and selecting appropriate methods of assay. Most of the methods used were existing biological and chemical tests, and are the ones presently in the USP monographs for the seven different insulin dosage forms currently tested.

About 1938, when the original patent was about to expire, FDA began certifying each batch of insulin produced, following the tests prescribed by 21 CFR 429 and the USP.

The basic source of new methods of insulin assay is the NDA, the counterpart of NCAA's Forms 5 and 6. All proposed test changes and new tests must be submitted by the manufacturer in its Master File or supplement. When the new or modified tests have been validated and accepted, they are submitted to the Federal Regulation Writers for inclusion into the CFR. In addition, they are submitted to the USP for inclusion in the next supplement or revision.

There are also semiofficial or "in-house" methods used by both industry and FDA. One is the radioimmunoassay developed for use on those products where the official biological tests are not suitable because of the type of manufacturing process used.

Many changes have been made in the final potency level or chemical form of the product in the more than 50 years since the discovery of insulin, but the major source of the hormone for human consumption remains the pancreases of slaughtered animals.

To summarize, NCAA and the Division of Drug Biology's Insulin Certification group utilize information from various sources including NDA's (Form 5), the sampling process, including certification and surveillance samples, the reprint collection, cross reference cards, and literature searches, as resources for developing their analytical methods.

## REFERENCES AND NOTES

(1) 21 CFR 431.50.
(2) M. Margosis, "GLC Determination of N,N-Dimethylaniline in Penicillins", *J. Pharm. Sci.*, 1634–6 (1977).
(3) G. G Carter, "A Review of Procedures for the Detection of Penicillin Residues in Drugs", FDA By-Lines No. 3, Washington, D.C., 1977, pp 119–137.
(4) D. C. Grove and W. A. Randall, "Assay Methods of Antibiotics. A Laboratory Manual", Medical Encyclopedia Inc., New York, N.Y., 1955, p 238.
(5) F. Banting and C. H. Best, *J. Lab. Clin. Med.*, 7, 251 (1921–2).

# The Development and Compilation of Chemical Information by the Bureau of Drugs Medical Library for the Drug Review Process†

ROY SOLKOT

Food and Drug Administration, Bureau of Drugs, Rockville, Maryland 20857

Received September 12, 1978

Supportive integration into FDA's drug review process of literature surveys of biochemical and biomedical material published worldwide may take place at any stage in the progress of the review. Technical information specialists on the Library staff utilize printed abstract/index sources as well as computer terminals on the Library premises to access at least 37 unique databases. Pertinence embraces numerous fields (pharmacology, toxicology, agriculture, biology, chemistry, food technology, human and veterinary medicine, patents, pesticides, pollution, and psychology), general scientific data, and both completed and projected or ongoing U.S. government-sponsored research.

Little reference as to how chemical information is developed and utilized by the Food and Drug Administration (FDA) Medical Library was made at the previous Symposium on Information Handling and Processing by the FDA presented before the Division of Chemical Information at the 1976 San Francisco Meeting.[1] I would like to introduce this rather fundamental feature into the current Symposium at this time.

The Library participates in all phases of the drug review process, and the chemical literature plays a major role in this support, such participation beginning with the development and organization of the chemistry collection and extending through the stages of retrieval and utilization of this information. The latter responsibility is our focus here.

## NATURE AND SCOPE OF LITERATURE SUPPORT

Literature searches support all drug program activities. These might include: initial IND/NDA review to substantiate, verify, or supplement submitted data; development of protocols, guidelines, and standards; background for advisory committee meetings; preparation for hearings; and review of prescription labeling and/or OTC drugs, etc. The bulk of literature searches is concentrated on such subject areas as bioavailability, pharmacokinetics, and pharmacology; analyses and physical properties; therapeutic and/or adverse effects, poisoning, and toxicity involving either single drugs or drugs in combination (including synergism and/or inhibition, in the latter case); carcinogenicity, mutagenicity, and teratogenicity; packaging, quality control, and storage; etc.

Many surveys must be carried out in part or entirely by manual retrieval since they require delving into the pre-1965 literature. They usually focus first on sources that provide abstracts and utilize, among others, such "hard copy" as *Biological Abstracts*, which contains an indexing technique in its format that results in a set of five indexes, any of which may be used alone or in combination with the others. Most important of these indexes, for our purpose, are (1) the natural-language Subject Index in which entries appear in an alphabetically permuted format (this was formerly termed BASIC (Biological Abstracts Subjects In Context) and was compiled from significant terms in the authors' own titles, as well as from added key terms from both bodies of abstracts and original articles, and (2) the Cumulative Concept Index (earlier called the CROSS (Computer Rearrangement Of Subject Specialties) Index) which facilitates retrieval according to approximately 580 subject concepts.

† Presented before the Division of Chemical Information, 175th National Meeting of the American Chemical Society, Anaheim, Calif., March 15, 1978.

As a sort of chemical counterpart of *Biological Abstracts*, *Chemical Abstracts* (CA) constitutes a second major source of information for manual retrieval. The individual-issue subject index (one per week) contains terms less formal than those preferred by CA and published separately in the cumulated subject index for each volume. That subject index, covering chemical substances and known as the Chemical Substance Index, contains terms mandated, along with their identifying CA Service Registry Numbers, in a separately published Index Guide, which usually includes the specific Registry Numbers assigned to those terms. To warrant entry in a CA Chemical Substance Index, a chemical compound must either be new or have new information reported on it. Known chemical compounds are indexed when useful data about them are reported in the original document.

The second of the two CA subject indexes comprises all those headings which do not refer to specific chemical substances and is designated General Subject Index. This index includes classes of chemical substances, incompletely defined materials, rocks (as distinct from specific minerals), physicochemical concepts and phenomena, reactions, engineering and industrial apparatus and processes, biochemical and biological subjects (other than specific biochemicals), and common and scientific names of animals and plants.

These two indexes are supplemented by (1) still others that specialize either in names of authors, molecular formulas, patents and their concordances, registry numbers, or index guides, and (2) collective indexes which cumulate ten CA volumes (five years). The 9th Collective Index, covering CA Volume 76 through 85 (1972–1976), is current at this time.

Other important abstract services that might be utilized in the manual mode are (1) the *Excerpta Medica* series, (2) *International Pharmaceutical Abstracts*, (3) *ISI Abstracts of Chemistry and Index Chemicus*, (4) *RINGDOC Abstracts*, and (5) our own, in-house-generated *Clinical Experience Abstracts*, with emphasis upon observed or possible adverse effects associated with the use of additives, cosmetics, devices, drugs, and household products—in humans only. The above-mentioned *RINGDOC Abstracts* service constitutes conventional summations of all the information contained in original documents and transferred to a heating of descriptors in uncoded English, in an abbreviated and standardized form. It includes abstracts of documents without quantitative details. Also representative of this source are De Haen Drugs in Use cards, with their information boxed under headings such as adverse reactions, class of drugs, concomitant therapy, diagnosis of disease treated, dosage used, laboratory findings, product results obtained, route of administration, sex and age of patients, etc.

The abstracts are supplemented wth index services calculated to retrieve information either (1) amenable to a bibliographical arrangement, like references from *Adverse Reactions Titles*, *Bioresearch Index*, *Current Contents*, *Drug Literature Index*, *Iowa Drug Information Services* (IDIS), *Index Medicus*, *Index Veterinarius*, and the like, or (2) in the form of card records of either trade names or research code numbers kept up to date and on file in the Library; keywords may be included.

Surveys that are carried out in whole or in part employing computerized database facilities introduce a choice of at least 37 chemically oriented and more or less unique databases directly accessible through computer terminals available on the Library premises. Pertinence includes such fields as pharmacology, toxicology, agriculture, biology, chemistry, food technology, human and veterinary medicine, patents, pesticides, pollution, and psychology, as well as general scientific data and both completed and projected or ongoing U.S. government-sponsored research.

Of the four corporate computer systems to which we have access, by contract, that provided by the Lockheed Information Systems "DIALOG" Service has almost 90 on-line databases, of which 24 appear to be pertinent to our chemical search pursuits; 14 of these are unique, not duplicated by any of the other three systems at our disposal.

Of these three systems, the System Development Corporation has about 36 on-line databases, of which 15 appear to be pertinent to our chemical needs; five of these are unique.

Of the 16 on-line databases provided by the Bibliographic Retrieval Services, four are pertinent and one is unique.

Out of a total of some 11 on-line databases provided by the National Library of Medicine, seven are pertinent and six unique.

In summation, then, we have about 50 pertinent databases, 26 of which are unique, five duplicated by three of the four systems, five more duplicated by two systems, and one duplicated by the other two systems, providing a total of 37 different on-line databases.

Several other retrieval systems not directly accessible on-line are available and include material from the IDIS, the Defense Documentation Center (DDC), and the Toxicology Information Response Center (TIRC), as well as a few more highly specialized federal sources from which may be obtained data relevant to general analytics, bacteriology, infrared, mass analysis, nuclear magnetic resonance, powder diffraction, etc.

Databases that focus upon material peripheral to chemistry are also frequently accessed. These include aquatic sciences and fisheries, business, ecology, engineering, environment, foundations directory, foundation grants index, grants of other kinds, geography, history, legislation, management, marketing, physics, sociology, statistics, textiles, etc.

The ability to retrieve relatively recent database entries on-line, while the operator sits at the computer-terminal console, provides prompt service to patrons. Most requestors are satisfied with receipt of off-line retrieval received via mail a few days after initiation of a request, and which covers older references, in some cases as far back as 1964. Most patrons are happy to be provided with complete citations and consider keywords and/or abstracts as a bonus, when these, also, are provided.

Continual updating of any subject, wherever necessary, is also available automatically, once initiated, through a Selective Dissemination of Information approach. This Library has four full-time technical information specialists available to handle searches.

Depending on requestors' needs, results (mostly in the form of references—often with accompanying abstracts and/or keywords or, in special cases, consisting of copies of complete technical articles or symposium abstracts, extracts from texts, etc.) may be compiled and arranged alphabetically (by first author, for example), chronologically, by subtopic, or in some other format.

With a view toward possible future use, copies of manually prepared bibliographies are subject indexed and kept on file in the Library. Searches consisting entirely of computer printouts are subject indexed as a unit on cards, but the printouts are not duplicated for retention in the Library for various reasons (e.g., the time-consuming feature associated with reproducing the often numerous entries in many of these printouts, the clerical support and storage space involved, and the rather poor photoreproducibility of some of the on-line printouts). We frequently borrow back printouts from users to lend to others with identical needs. However, by utilizing the "save search" feature built into some automated database systems, we can store a search strategy permanently and, on demand, retrieve a duplicate, including subsequent updates.

When the whole article from a title not in our collection or even a translation is desired, interlibrary loan, as well as contract and in-house translation facilities, are available.

At this writing, current information received from the SDC Search Service tells us that, among eight new or improved computer databases to be made available during the first quarter of 1978, there will be provided a Chemical Index from the CA Registry Nomenclature File and an initial loading of CHEM CASIA 10th Collective Index to date.[2] We also are looking forward to again using the *Excerpta Medica* on-line file when it becomes available through the Lockheed Information Systems "DIALOG" Service within a few months.

## REFERENCES AND NOTES

(1) S. W. Bellman, "Symposium on Information Handling and Poisoning by the Food and Drug Administration. Introductory Remarks", *J. Chem. Inf. Comput. Sci.*, **17**, 94–95 (1977); D. Banes, "USP and the Development of Drug Standards", *ibid.*, **17**, 95–96 (1977); W. Horwitz, "The Establishment of Official Analytical Methodology", *ibid.*, **17**, 97–102 (1977); H. P. Eiduson, "Application of Tolerances, Standards, and Methodology in the Enforcement of the Food, Drug, and Cosmetic Act", *ibid.*, **17**, 102–105 (1977); T. Cairns and R. A. Jacobson, "New Approaches to FDA Analytical Problems", *ibid.*, **17**, 105–109 (1977).
(2) "More New Orbit Data Bases", *Search Service News* 5(11), 1 (Dec 1977).

# DOVE, a Rational Analysis of Sparse Data

PETER F. STRONG

Arthur D. Little, Inc., Cambridge, Massachusetts 02140

C. GARDNER SWAIN* and MARGUERITE S. SWAIN

Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received August 28, 1978

DOVE is an unconventional statistical procedure which can be useful, even when many or most data are missing, for (1) generalized least-squares fitting to evaluate a self-consistent set of all parameters in an expression for predicting data, and (2) transforming the set of parameters obtained in phase 1, without changing the predicted data, so that each final parameter has a simple, pure, realistic, physical meaning. Since predicted data are expressed as $a_ix_j + b_iy_j + \ldots c_i$ with $n$ product terms, phase 2 requires incorporation of $n^2 + n$ independent subsidiary conditions, of which $2n$ are arbitrary, i.e., merely fix zero reference points and scale unit sizes, but $n^2-n$ are critical, i.e., must be relationships between particular parameters supported by other information. Both phases are illustrated by an example with $n = 2$, $1 \le i \le 7$, $1 \le j \le 10$, which therefore needs 41 parameters to fit 70 data plus the 6 subsidiary conditions. Valid parameters are obtained even if 30 of the 70 possible data are missing. DOVE provides a general method for evaluating all the parameters in linear free-energy relationships with two or more product terms.

## INTRODUCTION

DOVE is a handy procedure for predicting missing data and forcing every parameter in a fitted expression to have a simple, discrete, realistic, physical meaning. The acronym DOVE, standing for "dual obligate vector evaluation", refers to its two-phase evaluation of all parameters, obligating them by least squares in phase 1, and additionally obligating them in phase 2 by subsidiary conditions that are supported by information other than the data.

## METHOD

**Phase 1.** Equation 1 embodies the least-squares criterion of fit.[1]

$$\sum_i \sum_j e_{ij} w_i (z_{ij} - p_{ij})^2 = \text{minimum} \tag{1}$$

Here $z$ and $p$ refer to observed and predicted data, $j$ specifies the variable of main interest, $i$ specifies all other variables, $e_{ij}$ is unity if $z_{ij}$ exists but zero for any $ij$ combination not observed, and $w_i$ are suitable statistical weights.

Equation 2 is a generalized form of a widely applicable expression for predicted data.

$$p_{ij} = \sum_{m=1}^{n} s_{im} f_{mj} + c_i \tag{2}$$

Its parameters comprise factors $f$, slopes $s$, and intercepts $c$.[2] However, the confusion of double subscripts on factors and slopes can be avoided by a notation using different factor and slope symbols for each different product term or mode $m$. Therefore we will switch to expressions for $p_{ij}$ such as

$$p_{ij} = c_i \tag{3}$$

$$p_{ij} = a_i x_j + c_i \tag{4}$$

$$p_{ij} = a_i x_j + b_i y_j + c_i \tag{5}$$

$$p_{ij} = a_i x_j + b_i y_j + g_i q_j + c_i \tag{6}$$

as soon as we have decided on the number, $n$, of modes to include.

The subscripts $j$ and $i$ need elucidation. Subscript $j$ refers to the main or primary variable, while subscript $i$ refers to all other variables. To be more precise, each $j$ is a numerical index for one specific example of the principal variable. In the past, a specific example has been variously called a case, individual, object, entity, or unit.[3] Since most of these names are ambiguous or cumbersome, we will call it a "jot". Subscript $i$ is a numerical index for a group having a common set of all the other variables. This group has also been called a variable, attribute, characteristic, property, class, or series. We will call it an "ilk". For example, in a study of solvent effects the main variable is the solvent. A $j$ of 1 might denote that water is the jot, while a $j$ of 2 might identify the jot as ethyl alcohol. An $i$ of 1 might refer to an ilk composed of logs of rate

0095-2338/79/1619-0013$01.00/0 