Real-Time Interrogation of Chemical Data*

RUDOLPH J. MARCUS** and EUGENE E. GLOYE
Office of Naval Research,
1030 East Green St.,
Pasadena, Calif. 91106

Received April 1, 1971

A heuristic research methodology has been applied to narrative chemical and medical information taken from the "Merck Index." This approach is made possible by real-time interaction with a computer in which the scientist remains in the iterative loop. Techniques by which the narrative data are examined and manipulated are discussed. Problems concerning the structure and action of sympathomimetics, parasympathomimetics, mescaline, and indoles are elucidated by applying this research methodology.

Time-shared computer systems provide an environment within which the scientist can interact with his data in a heuristic manner. A definition of heuristic programming, taken from a recently published dictionary of computer terminology,[1] is:

> program, heuristic—1. A routine by which the computer attacks a problem not by a direct algorithmic procedure, but by a trial and error approach frequently involving the act of learning. 2. A set of computer instructions that simulates the behavior of a human operator in approaching a similar problem.

With real-time interaction, the heuristic procedures consist of a continued, iterative exchange between the man and the machine. According to Goodman, "The computer itself is being used for a variety of intelligent activities. In order of increasing difficulty, they are to sort, count, store and retrieve, perform mathematical operations, perform logical operations, and be creative."[2] For the computer to be truly creative, however, the scientist must continually be part of the iterative loop. This is possible with older, batch-processing methods, but is hindered and discouraged by lengthy job trains, unfavorable priority assignments, and consequent delays of hours, days, or even weeks. With a time-shared computer the interaction occurs in real time and is conducive to continued dialog. The machine thus shares with the scientist in a process very much like man's internal cognitive experience.

Instead of formulating a specific hypothesis and then gathering data that would validate or invalidate this hypothesis, there is an opportunity in the heuristic approach to begin with a comprehensive group of data and to examine it repeatedly and interactively with new hypotheses. Solomon,[3] among others, has described clustering techniques in which data manipulation produces both the number of groupings or clusters into which the data fall and the assignment of each datum or data element to these groupings. In fact, Solomon has argued for the use of numerical taxonomies derived in this manner in the place of more traditional data analysis techniques. This is an inductive approach which does not prescribe a priori the form in which the data should be organized. Such freedom of form is one reason why these methodologies may be less familiar to physical scientists than they are to behavioral scientists.

The availability of the high speed computer with its large data handling capacity at modest costs is the key to use of the heuristic technique with very large size data bases. Solomon concerns himself mainly with numeric data, but computer technology can extend the inductive approach to narrative (alphanumeric) data. For example, Eiduson and Ramsey-Klee,[4] writing about a large size data base containing narrative clinical material, say: "The use of the computer as a processor of narrative text opens an additional area of research methodology heretofore confined primarily to limited content analyses or experimental studies in verbal behavior."

One of the aims of the present work was to explore the feasibility of using the methodology referred to by Eiduson and Ramsey-Klee with narrative chemical and medical information. Its feasibility is demonstrated by examples reported in this paper. At the same time, some automatic coding procedures have been developed which will make it possible to perform numeric taxonomic analyses such as those described by Solomon.

THE MERCK INDEX DATA BASE

The chemical and medical information to which this new methodology was applied consisted of the chemical names and medical uses of compounds listed in the "Merck Index."[5] All compounds for which a medical use is given (about 1000 in the 6th edition, 1952, and about 4000 in the 8th edition, 1968) were included. All text printed under the subheading "Medical Use" was copied verbatim. Chemical information copied included the common name, all synonyms, and the empirical formula.

Four separate files were created to facilitate the retrieval of different kinds of information. Two of these consist of medical use entries from the 8th and 6th editions of the Index, respectively, the third contains the empirical formulae, and the fourth (largest) file lists all of the chemical synonym names (trade names etc.).

Each line of the files consists of two variable-length fields.

*Presented at the symposium on "Input-Output Interaction of the Chemist with the Computer," 161st Meeting, ACS, Los Angeles, Calif., March 31, 1971.

**To whom correspondence should be addressed.

The two fields are separated by a unique character ("≠") which does not appear in the Index text and which is not used in the programming software. The first of the two variable-length fields in each line is occupied by the common name of the compound. The other field in the line is occupied by narrative information, with the text continuing to the second field of succeeding lines when necessary. When the character string in the second field extends to more than one line per compound, the first field of each line contains the common name or, if that is too long, some abbreviation after the common name is once spelled out in full. (Only one common name among the 4000 compounds with medical uses found in the 8th edition of the "Merck Index" was longer than 72 characters, the line length for standard teletypewriter-formatted text.) The common name thus serves as a tag for each medical use, synonym, or empirical formula.

The compound name and associated text constitute an entry in each of the four files. The four entries for the compound benzthiazide appear as follows in the "Merck Index" data base:[6]

Medical Use File, 8th Edition

```
BENZTHIAZIDE ≠ DIURETIC, ANTIHYPERTENSIVE, DOSE:  ORAL 25 TO 50 MG.
BENZTHIAZIDE ≠   SIDE EFFECTS:  ANOREXIA, NAUSEA, DIZZINESS, HEADACHE
BENZTHIAZIDE =     MAY OCCUR.
```

Medical Use File, 6th edition

(does not appear in 1952 edition because preparation was first reported in 1959)

Empirical Formula File

```
BENZTHIAZIDE ≠ C15H14CLN3O4S3.
```

Synonym File

```
BENZTHIAZIDE ≠  3-((BENZYLTHIO)METHYL)-6-CHLORO-2H-1,2,4-BENZOTHIADIAZINE-
BENZTHIAZIDE ≠   7-SULFONAMIDE 1,1-DIOXIDE; 3-((BENZYLTHIO)-METHYL)-
BENZTHIAZIDE =     6-CHLORO-7-SULFAMOYL-2H-BENZO-1,2,4-THIADIAZINE
BENZTHIAZIDE =      1,1-DIOXIDE; 6-CHLORO-7-SULFA-MOYL-3-BENZYLTHIOMETHYL-
BENZTHIAZIDE =       2H-1,2,4-BENZOTHIADIAZINE 1,1-DIOXIDE; URESE;
BENZTHIAZIDE =        BENZOTHIAZIDE; AQUATAG; FREE URIL; EXOSALT; EXNA.
```

(The transliteration rules followed in going from natural text to machine-readable text will be apparent upon comparison of the above three entries with the paragraph "Benzthiazide" as it appears on p. 137 in the 8th edition of the "Merck Index.")

COMPUTER MANIPULATION OF NATURAL TEXT

Commercial time-shared computer services generally offer their customers software packages with an array of string manipulation capabilities. The original purpose of interactive string manipulation and editing was to facilitate the on-line debugging of batch programs. Experience has shown that the same software is efficient and appropriate for retrieving and clustering narrative data. For example, in the "Editor" language offered by Tymshare, Inc., there is a find command ("FIND") which allows the user to retrieve any specified alphanumeric string from files to which he has access. Lines containing the specified string are printed out at the teletypewriter terminal in real time. At the option of the user, the retrieved lines may be segregated in a new data file for later manipulation, which may include a printout in batch mode on a high-speed printer.

In addition to extensive use of the find command, this work has also relied heavily upon a library sort-merge program adaptable to both fixed- and variable-length fields. (This program, written in "SuperBasic," is called "Supersort" in the user library of Tymshare, Inc., 336 East Kelso St., Inglewood, Calif. 90301. This program is self-instructing when called.) Concatenation dictated by limitations on file size is accomplished interactively in the "Editor" language. Other uses of the "SuperBasic" language for manipulation of natural text are described in the last section of this paper.

Several cases will now be examined to illustrate the interactive use of available software. A question arose about the various medical uses of compounds containing the indole ring. Here the search string used was "indo." This string was chosen so that indo-substituted compounds as well as those named as substituted indoles would be recovered. The search was made in the synonym file. The following is a verbatim reproduction of the user-formulated command and of a 3-compound sample of the computer output. Note that the underlined search string typically occurs as part of the Geneva system name in the second field of each line in the synonym file. (Use of the Geneva system of nomenclature for storing, retrieving, and sorting chemical structure information in machine-readable form has been referred to previously.[6] The primary use of structural information in this work is as one of the hyperspace coordinates in clustering, rather than as the basis for bibliographic retrieval.)

```
FIND "INDO"/
ADRENOCHROME ≠  3-HYDROXY-1-METHYL-5,6- INDOLINEDIONE.
               :
BUFOTENINE =  3-(2-DIMETHYLAMINOETHYL)INDOL-5 OL;

BUFOTENINE =   3-(2-DEMETHYLAMINOETHYL)-5-INDOLOL; 5-HYDROXY-

BUFOTENINE =      5-HYDROXYINDOLE; MAPPINE.
              :
TYLOCREBRINE =      2,3,5,6-TETRAMETHOXYPHENAN-THRO(9,10:6',7')INDOLIZIDINE.
```

(The purpose of successive indentations of the second field in multi-line entries becomes clear in this example, where the third line of the four-line entry "Bufotenine" did not print out because it did not contain the search string "indo.") Of the 27 compounds found in this search, the medical use file shows that 19 have adrenergic effects, and the others do not. The chemical difference between these two groups of compounds containing indole rings is that the ones which show adrenergic effects are so substituted that they can undergo quinone methide formation and can thus act as adrenergic agonists.[7]

Another question which the data base was asked concerned the mode of action of mescaline. Only a single compound is recovered from the synonym file when the search string "mescaline" is used, showing that the Index does not name close chemical relatives of mescaline as mescaline derivatives. To find some compounds with similar chemical structure, the search string "phenethylamine" was used in a find command. The following is a verbatim reproduction of the user-formulated command and of a 3-compound sample of the computer output. (As an aid to the reader's understanding of how the search command functions, the search string is underlined in each of the lines of output.)

```
FIND "PHENETHYLAMINE" /
AMPHECLORAL ≠ ALPHA-METHYL-N-(2,2,2-TRICHLOROETHYLIDENE)-PHENETHYLAMINE;
           :
MESCALINE ≠ 3,4,5-TRIMETHOXYPHENETHYLAMINE; MEZCALINE.
           :
XYLOPROPAMINE ≠ ALPHA,3,4,-TRIMETHYLPHENETHYLAMINE;
```

With this search command, 28 compounds were recovered. Consultation of the medical use file with individual find commands resulted in the following computer output. (Consecutive use of two separate files such as the synonym file and the medical use file becomes unnecessary when the coding procedures described in the last section of this paper are used.)

```
FIND "AMPHECLORAL"/
AMPHECLORAL ≠ SYMPATHOMIMETIC, ANOREXIGENIC.
FIND "MESCALINE"/
MESCALINE ≠ EXPTL PSYCHOTOMIMETIC AGENT.
FIND "XYLOPROPAMINE"/
XYLOPROPAMINE ≠ SYMPATHOMIMETIC. HAS ANALGESIC PROPERTIES.
```

The 28 compounds included various amphetamines and all but mescaline are described as sympathomimetics. Because mescaline clusters with the other compounds as a phenethylamine, one concludes that mescaline may exercise its "psychotomimetic" action by activating the sympathetic nervous system just as do the other compounds in the cluster.

COMPUTER-AIDED CLUSTERING

In the previous examples, chemical names or parts of chemical names formed search strings. The data base can be entered equally well through medical use terms of interest. It will be recalled that the medical use information was copied in unedited narrative form to construct the medical use files. This text always appears in the second field of each line.

A topic of both chemical and behavioral significance is the structure of compounds which act on the autonomic nervous system. (The time-shared computer environment has made it possible for the authors, a chemist and a psychologist, to bridge the wide conceptual gap between these disciplines.) Investigation of this topic presents an appropriate example of how medical use terms are used in find commands. The search string "sympatho" produced 73 compounds, of which 50 are sympathomimetics and 23 are parasympathomimetics. These compounds are listed in Table I.

Sympathomimetics are compounds which cause the organism to act as if its sympathetic nervous system had been stimulated.[8] These compounds have an ethylamine structure. Among these ethylamines are catecholamines like epinephrine, norepinephreine, dopamine, and serotonin, which are known to be nerve impulse transmitters in the sympathetic nervous system. It appears from the printout that absorption of other ethylamines results in arousal, similar to the arousal caused by release of catecholamines which are naturally present in the organism. Subsequent consultation of the pharmacological literature[9] and discussion with researchers active in the field[10] demonstrated that indeed this is the case. Increased degrees of arousal which are caused by substituted ethylamines such as mescaline and the amphetamines are well-known phenomena of this kind.

Parasympathomimetics are compounds which cause the organism to act as if its parasympathetic nervous system had been stimulated.[8] These compounds are substituted ammonium ions. Among those organic ammonium ions is acetylcholine, which is known to be a nerve impulse transmitter in the parasympathetic nervous system. It appears from the printout that absorption of other organic ammonium compounds or, going to the next row of the periodic table, of organic phosphates affects the regulation of autonomic functions as release of naturally occurring acetylcholine does. Subsequent consultation of the pharmacological literature demonstrated that indeed this is the case.[11] Regulation of the autonomic functions is a delicate matter, and many of the parasympathomimetics are highly poisonous. The phosphates, in particular, are used as insecticides.

Tables which list sympathomimetic and parasympathomimetic compounds can be found in modern pharmacology textbooks.[9, 11, 12] The heuristic methodology underlying this work gains validity through comparison of computer outputs with independent sources in the contemporary literature.

The "Merck Index" does not use the terms sympatholytic and parasympatholytic which denote the action opposite to sympathomimetics and parasympathomimetics.

Table I. Compounds Described as Sympathomimetics and Parasympathomimetics in the "Merck Index"

Sympathomimetics		Parasympathomimetics
Isoproterenol	Methoxyphenamine	Arecoline
Protokylol	Methylhexaneamine	Carbachol
Xylometazoline	Hydroxyamphetamine	Bethanechol Chloride
KB 227	Ethylphenylephrine	Benzpyrinium Bromide
Naphazoline Hydrochloride	2,2'-Dicyclohexyl-*N*-methyl-diethylamine Hydrochloride	Acetylcholine Bromide
Tetrahydrozoline	*N*,1-Dimethylhexylamine	Acetylcholine Chloride
Tramazoline	Dioxethedrine	Pyridostigmine Bromide
Oxymetazoline	Etafedrine	Oxapropanium Iodide
Phenylpropylmethylamine	Homarylamine Hydrochloride	Neostigmine Bromide
Tuaminoheptane	Ethylnorepinephrine	Neostigmine Methyl Sulfate
Propylhexedrine	Hordenine	Methacholine Bromide
Phenylephrine Hydrochloride	*p*-Hydroxyephedrine	Methacholine Chloride
Nordefrin Hydrochloride	Octopamine	Hexadistigmine
Isometheptene	Phenmetrazine	Geneserine
Ephedrine	Synephrine	Edrophonium Bromide
Epinephrine	Tyramine	Demecarium Bromide
Deoxyepinephrine	Xylopropamine	Pilocarpine
Phenylpropanolamine Hydrochloride	Benzphetamine	Pilocarpine Hydrochloride
Cyclopentamine	Cyclexedrine	Pilocarpine Nitrate
α-(Aminomethyl)-*m*-hydroxybenzyl Alcohol	Dextroamphetamine Sulfate	Diisopropyl Fluorophosphate
Pholedrine	Tanphetamin	
Norepinephrine	Amphetamine	Diethyl *p*-Nitrophenyl Phosphate
Metaraminol	Amphetamine Phosphate	Diisopropyl *p*-Nitrophenyl Phosphate
Mephentermine	Amphetamine Sulfate	Tetraethyl Pyrophosphate
Methoxamine Hydrochloride	*d*-Desoxyephedrine Hydrochloride	

Therefore, chemical structure information was used to recover those compounds. A search for ethylamines in the synonym file produced not only those compounds which are sympathomimetics, but also a number of compounds which are sympatholytics. The structural difference between sympathomimetics and sympatholytics appears to be that sympatholytics are ethylamines in which the amine itself is protected or hindered. An ethylamine substituted with bulky side groups might block the action of naturally occurring catecholamines by occupying a site meant for these nerve impulse transmitters. Absorption of such a compound would be expected to have a calming effect because nerve impulses characteristic of arousal would be transmitted less often by the catecholamines. The reserpine alkaloids are ethylamines which contain bulky side groups, and are well-known tranquilizers.

A similar search for organic ammonium ions in the synonym file produced not only parasympathomimetics, but also a number of parasympatholytics. The structural difference between parasympathomimetics and parasympatholytics again appears to be that the characteristic ammonium ion structure is present in parasympatholytics, but is obstructed or hindered. An organic ammonium ion substituted with bulky side groups might block the action of acetylcholine by occupying a site meant for this nerve impulse transmitter. Tubocurarine chloride is an organic ammonium ion which contains bulky side groups. It is known to function as a skeletal muscle relaxant by blocking acetylcholine.[13] It can even cause respiratory paralysis. Other organic ammonium ions with bulky side groups are curarimimetic agents which exercise a similar action. These include hexafluorenium bromide, dipropamine, echitamine, beta-erythroidine, dihydro-beta-erythroidine, laudexium methyl sulfate, gallamine triethiodide, and hexacarbacholine bromide.

This example shows how computer-aided clustering has been used in a narrative-style data base to identify molecular substructures which appear in all sympathomimetics (ethylamines) and in all parasympathomimetics (ammonium ions). These molecular substructures appear in the naturally occurring nerve impulse transmitters (catecholamines in the sympathetic nervous system, acetylcholine in the parasympathetic nervous system) and in those compounds which mimic their action. When these same molecular substructures which appear in all sympathomimetics they become antagonists of the naturally occurring nerve impulse transmitters (reserpine alkaloids in the sympathetic nervous system, curare in the parasympathetic nervous system). These relationships are presented in tabular form in Table II.

INVERSION OF THE "MERCK INDEX"

To run clustering programs efficiently, it became desir-

Table II. Schematic of Structure—Action Relation Among Compounds Affecting the Autonomic Nervous System

STRUCTURE	PHENETHYLAMINE	AMMONIUM ION OR PHOSPHATES
SYSTEM	SYMPATHETIC	PARASYMPATHETIC
TRANSMITTER	CATECHOLAMINES	ACETYLCHOLINE
HINDERED STRUCTURE	–LYTIC	–LYTIC
UNHINDERED STRUCTURE	–MIMETIC	–MIMETIC

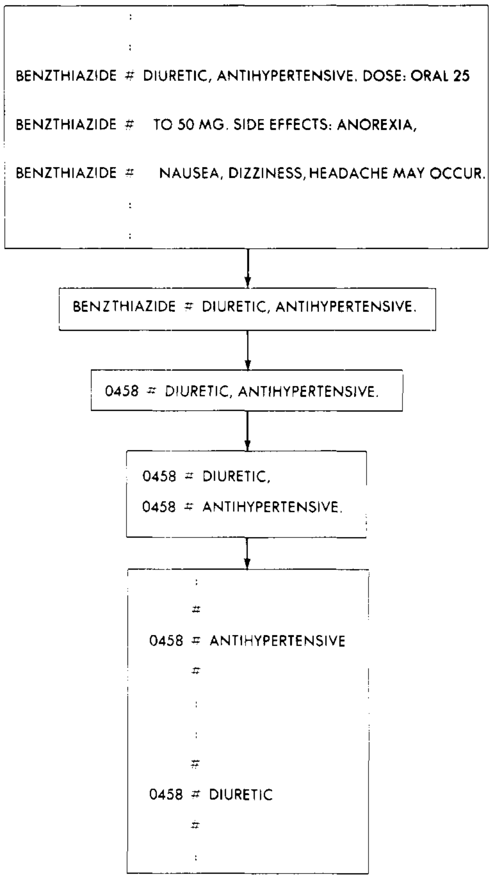
BENZTHIAZIDE ≠ DIURETIC, ANTIHYPERTENSIVE. DOSE: ORAL 25
BENZTHIAZIDE ≠ TO 50 MG. SIDE EFFECTS: ANOREXIA,
BENZTHIAZIDE ≠ NAUSEA, DIZZINESS, HEADACHE MAY OCCUR.
BENZTHIAZIDE ≠ DIURETIC, ANTIHYPERTENSIVE.
0458 ≠ DIURETIC, ANTIHYPERTENSIVE.
0458 ≠ DIURETIC,
0458 ≠ ANTIHYPERTENSIVE.
0458 ≠ ANTIHYPERTENSIVE
0458 ≠ DIURETIC

Figure 1. Illustration of algorithm used to invert "Merck Index" entries

able to code both chemical compounds and medical uses with a letter or number code. This coding was done in several steps using interactive computer techniques. The algorithm employed is illustrated in Figure 1. In brief, the algorithm involved separating an initial 256-character line from the narrative entries, numbering each of these lines serially (which gave each compound a unique number code), and dropping all text past the first period. Because individual medical uses are set off by commas in the "Merck Index," the computer was then instructed (using, throughout, programs written in "SuperBasic") to create a new line for each phrase within the first sentence. Each of these phrases, now containing only one medical use (often only one word) and the number code of the compound associated with that use, was then sent through a sort-merge program, alphabetizing on the use string, thereby completing an inversion from compound indexing to medical use indexing. (The file containing 4000 compounds arranged alphabetically by individual medical use is now being prepared for hard-copy printout.) Notice that in the manipulation of natural narrative data, syntax gives clues for meaningful separation rules.

By operating in an interactive mode, programs were checked on small samples of data before general runs were initiated. Frequently, minor inconsistencies in the text proved a hindrance in the computer processing. The task would have been extremely difficult to accomplish in the batch mode with characteristic waits for program debugging and program modifications. By working on-line and by inspecting output as the computer processed the narrative data, the processing proved to be moderately easy. At various stages in the task, the machine was employed to correct inconsistencies in the text to ensure appropriate processing. While text editing proved a laborious task, even with the aid of the interactive "Editor" language employed, it would have been so time-consuming as to be completely unfeasible as a hand operation without the computer.

The manipulation of natural narrative text depends on the syntactical features of language. Syntactical rules which are generally observed in the preparation of text make it possible to manipulate this information without previous specification of data fields or records. Obviously, greater efficiency can be realized if fixed fields of known contents are employed in organizing data for computer manipulation. However, when one begins a research effort within which the nature of the data is not clear because the data are too massive in scope to picture them accurately without automatic data processing, one must depend upon the natural syntactical features at the price of some machine inefficiency.

ACKNOWLEDGMENT

The authors wish to acknowledge with thanks the support provided by H. W. Fox, Director of ONR Chemistry Programs, Marvin Denicoff, Director of ONR Information Systems Programs, and G. C. Tolhurst, formerly Director of ONR Physiological Psychology Programs. We also wish to acknowledge the generous advice and help given by E. T. Florance of this office.

LITERATURE CITED

(1) Sippl, C. J., "Computer Dictionary," p. 144, Bobbs-Merrill, New York, N. Y., 1970.

(2) Goodman, A. F., "The Interface of Computer Science and Statistics," *Naval Logistics Research Quarterly*, June 1971, in press.

(3) Solomon, Herbert, "Numerical Taxonomy," Tech. Rept. No. **167**, Stanford University Department of Statistics, Stanford, Calif., December 1970.

(4) Eiduson, B. T., and D. M. Ramsey-Klee, "A Strategy for Life History Research Using Computer-Based Information Processing Techniques," in "Life History Research in Psychotherapy," M. Roff and D. F. Ricks, Eds., University of Minnesota Press, Minneapolis, Minn., 1970.

(5) Stecher, P. G., "The Merck Index," Merck, Rahway, N. J., 6th ed., 1952; 8th ed., 1968.

(6) Gloye, E. E., and R. J. Marcus, "Drug Effect Prediction by Computer," *Science* **169,** 89–91 (1970).

(7) Larsen, A. A., "Catecholamine Chemical Species at the Adrenergic Receptors," *Nature* **224,** 25–7 (1969).

(8) Patton, H. D., "The Autonomic Nervous System," in "Medical Physiology and Biophysics," 18th ed., pp. 220–33, T. C. Ruch and J. F. Fulton, Eds., Saunders, Philadelphia, Pa., 1960.

(9) Cutting, W. C. "Sympathetic Stimulants or Adrenergic Agents," in "Handbook of Pharmacology," pp. 499–518, Meredith, New York, N. Y., 1969.

(10) Barchas, J. D., Stanford University Medical School, Stanford, Calif., private communication.

(11) Cutting, W. C. "Parasympathetic Stimulants or Cholinergic Agents," in "Handbook of Pharmacology," pp. 528–37, Meredith, New York, N. Y., 1969.

(12) Innes, I. R., and Mark Nickerson, "Drugs Acting on Postganglionic Adrenergic Nerve Endings and Structures Innervated by Them (Sympathomimetic Drugs)," *in* "The Pharmacological Basis of Therapeutics," L. S. Goodman and Alfred Gilman, Eds. 4th ed., p. 485, Macmillan, New York, N. Y., 1970.

(13) Miller, N. E., "Learning of Visceral and Glandular Responses," *Science* **163,** 434–45 (1969).