NIOSH TECHNICAL REPORT

THE DEVELOPMENT AND APPLICATION OF ALGORITHMS FOR GENERATING ESTIMATES OF TOXICITY FOR THE NOHS DATA BASE

HERBERT L. VENABLE

U.S DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control
National Institute for Occupational Safety and Health
Division of Surveillance, Hazard Evaluations and Field Studies
Cincinnati, Ohio 45226

July, 1986

DISCLAIMER

Mention of company names or products does not constitute endorsement by the National Institute for Occupational Safety and Health

ACKNOWLEDGEMENT

The development and application of the algorithms presented in this technical report were accomplished under NIOSH contracts 210-78-0077, 210-78-0066, and 210-80-0044 (Genesee Computer Center, Inc., with Health Designs, Inc. and the Franklin Research Institute under subcontract).

I would like to thank the following people for their critical review of this document:

Ms. Alice Griefe
Occupational Toxicologist
Industrial Hygiene Section
Industrywide Studies Branch
DSHEFS, NIOSH

Dr. Harold Resnick
Science Advisor
Office of the Director
DRDS, NIOSH

Dr. Sanford Leffingwell
Chief, Research Analysis Section
Priorities Research Analysis Branch
DBBS, NIOSH

Dr. Curtis Travis
Office of Risk Analysis
Health and Safety Research Division
Oak Ridge National Laboratory

Dr. Robert W. Mason
Technical Advisor for Science
DBBS, NIOSH

Dr. Joseph Kelaghan
Epidemic Intelligence Service Officer
National Cancer Institute

I would also like to express my thanks to Dr. Wm. Karl Sieber, Jr. of NIOSH for his review of the statistical methodologies, Mr. David H. Pedersen of NIOSH for his suggestions and comments on organizing and writing this document, and to Ms. Kathy Mitchell for manuscript preparation.

DHHS (NIOSH) Publication No. 87-101

TABLE OF CONTENTS

TABLES

FIGURES

APPENDICES

* These appendices have been placed on microfiche and attached to the back cover of the printed report.

Special Note from the Author

The research and final products of the work presented in this report were accomplished under NIOSH contracts 210-78-0077, 210-79-0066, and 210-80-0044, with the author serving as the NIOSH Project Officer. However, time and funds allocated for this project expired before an approved final report was submitted by the contractor.

Since the author believes that the products of this project have significant value in the field of occupational safety and health, results of the project are reported here despite the lack of a final report from the original contractor.

The author wishes to extend his appreciation to and acknowledge the following individuals for their contribution, through the draft report, in the compilation of this report:

Mr. Kurt Enslein, President
Health Designs, Inc.
Rochester, New York

Dr. Paul Craig
National Library of Medicine
Bethesda, Maryland

Dr. John Strange
Franklin Research Institute
Philadelphia, Pennsylvania

Mr. Tom Lander
Health Designs, Inc.
Rochester, New York

Mr. Michael Tomb
Health Designs, Inc.
Rochester, New York

The text of this report is extracted largely from the contractor's incomplete draft report and is cited extensively throughout this report as are several publications by Enslein et al, which were written and published as the models were developed. These publications should be consulted in conjunction with this report to obtain a more comprehensive understanding of the project and its intent. Copies of the contractor's incomplete final report to NIOSH are available upon request from the author.

ABSTRACT

This project developed computer-based algorithms designed to provide estimates of toxicity for four toxicologic endpoints; LD_{50} (oral, rat), mutagenicity, carcinogenicity, and teratogenicity. These algorithms are the end result of a series of models tested against available toxicity data for each of the four toxic endpoints. The modeling data base for each endpoint contained a listing of chemical compounds determined to be toxic or non-toxic for each endpoint based on a subjective analysis of the bioassay data available.

Once the algorithms had been developed and tested, they were applied to the chemicals in the National Occupational Hazard Survey (NOHS) data base to generate estimates of toxicity for those chemical compounds known to be in the workplace. These estimates of toxicity are particularly useful in assessing the toxicity of those chemical compounds for which little or no toxicity data has been reported.

The algorithms produce estimates of toxic effect based on statistical computation and are therefore known to incorporate a certain degree of unavoidable statistical error. This and other limitations discussed in the report preclude the use of such theoretical toxicity data as a substitute for reported animal bioassay data or as the sole basis in making regulatory or other decisions of similar magnitude regarding the use of and exposure to chemical compounds. Instead, these toxicity data are intended only for rank-ordering a list of compounds according to relative toxicity or as a part of an overall process of selecting, testing, and evaluating chemical compounds for toxicity.

I. Introduction

A. Purpose

This project developed and applied computer-based algorithms to the chemical compounds (hereafter referred to as compounds) listed in the NIOSH National Occupational Hazard Survey (NOHS) data base in order to generate estimates of toxicity for these compounds for the following toxic endpoints:

LD_{50} (oral, rat)
Mutagenicity
Carcinogenicty
Teratogenicity

The theoretical toxicity data thus generated is intended for use only as an additional tool in assessing the toxicity of those compounds found in the workplace.

The compounds listed in the NOHS data base are a result of the National Occupational Hazard Survey which was a two-year study (1971-74) "intended to describe the health and safety conditions in the American work environment and, more specifically, to determine the extent of worker exposure to chemical and physical agents" (1). Observational data were gathered by surveying approximately 5,000 facilities encompassing all types of industrial activity covered by the Occupational Safety and Health Act (OSHA) of 1970. Approximately 8,000 separate chemical substances were identified as present in the workplace during the course of the survey. These 8,000 plus chemical substances are included in the NOHS data base.

The application of these four algorithmns to these compounds known to be in the work environment extends the utility of the data base by providing NIOSH with a unique toxicology information resource. Such a resource can be effectively utilized in a number of areas. For NIOSH, a major application could be for risk assessment and prioritization of research on chemical hazards in the workplace.

B. Structural Activity Relationships (SARs)

All four algorithms were developed on the assumption that a structure-activity relationship (SAR) exists among groups of compounds that exhibit similar chemical characteristics. For example, a SAR may exist among a group of compounds that possess a certain degree of ionic charge per molecule and may therefore have a similar degree of water solubility. SARs may be based on one or more of a number of molecular structure descriptors. Some of the more commonly used structural parameters are listed in Table 1.

The concept of SARs has been applied in several areas. For example, the primary use of the SAR concept in pharmaceutical chemistry has been for the evaluation of therapeutic effects of potential new drug compounds. Several approaches have been used in the application of

TABLE 1. A SAMPLING OF MOLECULAR DESCRIPTORS USED IN STRUCTURAL ACTIVITY RELATIONSHIP STUDIES

Physiochemical descriptors

Molecular weight
Density
Melting point
Boiling point
Logarithm of n-octyl alcohol/
water partition coefficient
Molecular refractivity*

Topological descriptors

Atom and bond fragments
Substructures (atom groups)
Substructure environment
Number of carbon atoms
Number of rings (in polycyclic compounds)
Molecular connectivity (extent of branching)

Geometrical descriptors

Molecular volume
Molecular shape
Molecular surface area
Substructure shape
Taft steric parameter*
Verloop sterimol constants*

Electronic descriptors

Hammett-Taft sigma constants*
Electron density -- bond reactivity
Dielectric constant
Dipole and higher moments
Ionization potential
Electron affinity

* These "complex descriptors" could be placed in other categories as well.

Reprinted with permission from Chemical and Engineering News, March 9, 1981 (2).

SAR research. Craig and Enslein (3) divided these methods of approach into four categories.

1. Intuitive Approach - which applies the organic chemists' skill, knowledge, and intuition. More recently this approach has focused on creating an additive model SAR which is based on the hypothesis that each structural feature of a molecule plays a consistent role in contributing to the overall activity of the molecule.

2. Multiple Parameter Approach - which combines known physical-organic chemical relationships into a novel mathematical expression to relate the biological activities of a closely related series of compounds to one or more physical properties (e.g., water-octanol solubility ratio or more commonly referred to as the partition coefficient).

3. Quantum Chemical Approach - which employs the principles of quantum mechanics and calculations. For example, one approach obtains electronic indices for a series of structurally related chemicals.

4. Substructural Analysis Approach - which is based on the analysis of type and, in some cases, frequency of occurrence of substructural or molecular fragments of molecular substructures, (e.g.,$-NO_2$).

Unlike the multiple parameter, additive model, or the intuitive approach methods, Adamson et al, state that the substructural analysis method may be used for a large number of structurally well-diversified compounds (4). Statistical analysis may then be applied to the type and frequency of substructural fragments to provide a quantitative value (i.e., coefficient value) for specific fragments that represents the amount of influence that each fragment exerts in the overall statistical variation of a group of compounds.

II. Development of Algorithms

A. General Background

Prior to 1975, the concept of SARs was generally applied to groups of structurally similar compounds, usually for the purpose of evaluating potential therapeutic effects in new drug research. Beginning about 1975, SAR concepts were applied to structurally similar groups of compounds for evaluating toxicity (5-10). Papers presented at the Symposium on Structural Correlates of Carcinogenesis and Mutagenesis, held at the U. S. Naval Academy, Annapolis, Maryland, 1977, reflect some of the areas of interest, endeavor, and success in application of SAR concepts for the evaluation of toxicity (11).

The application of quantitative structure-activity relationships (QSARs) to structurally diverse compounds for the evaluation of toxicity was first reported by Craig and Waite (12) and Enslein and

Craig (13). This project is an extension of this application of SAR concepts and employs the substructural analysis approach described by Enslein et al, (3).

B. Modeling the Algorithms

A number of molecular descriptors were considered for use in modeling the algorithms, (e.g., octanol-water partition coefficients and molar connectivity indices). In this project, regression analysis was used to select those molecular descriptor parameters most useful in modeling the algorithms. Ultimately, the occurrence of substructural fragments (and, in the carcinogen and LD_{50} models, molecular weight) were selected and used as the chemical descriptor variables in these algorithms.

All four algorithms were developed in a similar fashion. However, there were some differences and these will be pointed out in the presentation of the individual models. Basically, the procedure was as shown in Figure 1. A data base was created for use in developing each model. These data bases listed compounds selected on the basis of evidence indicating their ability to induce or not induce the effect of the selected toxicologic endpoint (e.g., carcinogen or noncarcinogen). Once the modeling data base was established, the resulting algorithm was designed and tested and then applied to the compounds listed in the NOHS data base for which the required information, (molecular formula, molecular weight, and a Wiswesser Line-Formula Notation) was available or could be generated.

Molecular structure plays a key role in all four algorithms in that a multi-step process is used to translate molecular data from a three-dimensional concept to a quantifiable value useful in generating toxicity estimates. These steps are summarized in Figure 2.

Wiswesser Line-Formula Notation (WLN) is used as the initial step in this translation process. The use of WLN is summarized by Smith and Baker as "...a precise and concise means of expressing the structural formulas of chemical compounds. Its basic idea is to use letter symbols to denote functional groups (chemical) and to use numbers to express the lengths of alkyl chains and sizes of rings. These symbols then are cited in connecting order from one end of the molecule to the other" (14)

The symbols employed by the WLN are the numerals 1-10, the 26 capital letters, the four punctuation marks &, -, /, and *, and a blank space (See Appendix A). According to Smith and Baker (14), with these symbols and approximately "a dozen new chemical symbols to supplement the old familiar ones, plus half a dozen operating symbols and the fundamental rules for manipulating them", a chemist should be able to write a WLN or read one as you would read a conventional structural formula.

As might be expected, the accuracy and usefulness of a toxic endpoint prediction, as estimated by these four algorithms, depends largely on an accurate description of the molecular structure.

FIGURE 1. PROCEDURES FOR DEVELOPING AN ALGORITHM

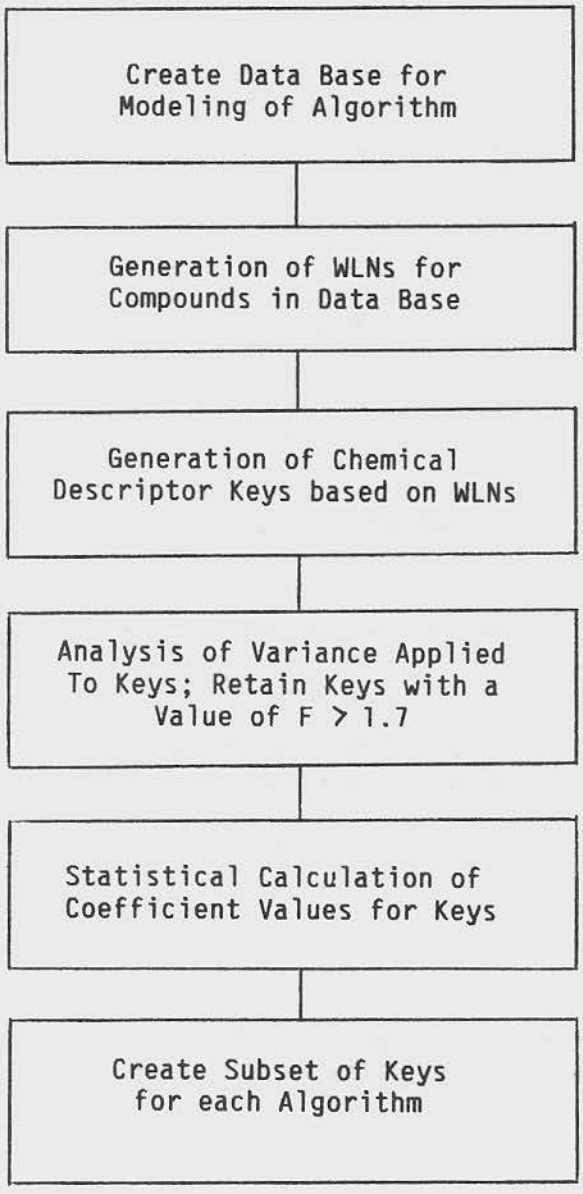
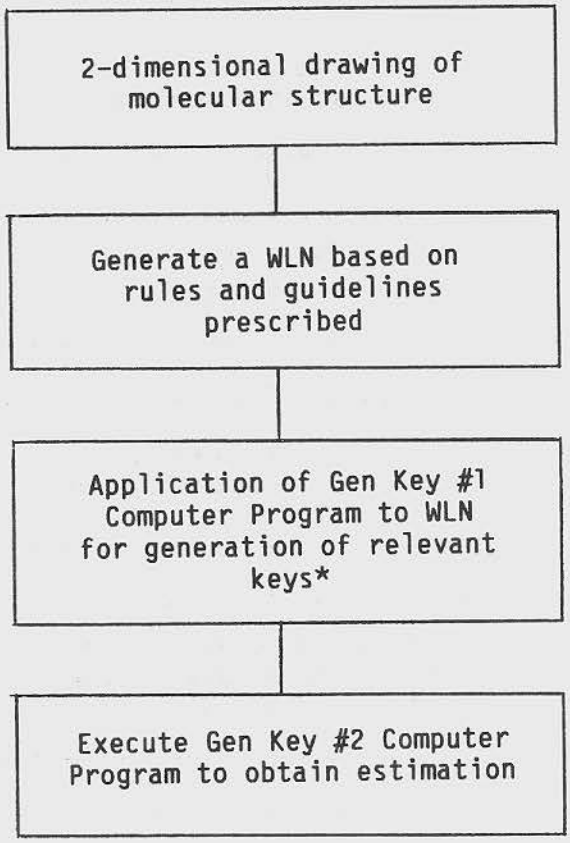


FIGURE 2. TRANSLATION PROCESS FOR OBTAINING A QUANTIFIABLE (NUMERICAL) REPRESENTATION OF MOLECULAR STRUCTURE



* Note that keys listed as problem keys (Table 2) must be manually checked as being relevant or not to the assigned WLN.

The accuracy of the generated WLN is subsequently expressed in the generation of the relevant chemical descriptor keys which are numerical representations of substructural molecular fragments. For example, the -OH (hydroxyl) substructure is the letter Q in WLN and Key 38 in chemical descriptor key terminology.

Assigning an accurate WLN to a compound requires a complete knowledge of Wiswesser Line-Formula Notation in conjunction with a considerable background in organic chemistry structure and nomenclature. However, techniques have been developed for generating WLNs by drawing the structures on an electronic graphics pad linked to an appropriately programmed computer which then generates the WLN (15).

A WLN cannot be accurately generated for certain compounds. Most notable of these are polymers or compounds for which the molecular structure may vary or is not known. It was also determined that inorganic compounds do not perform well in any of the four models (3). This is due largely to the inadequate WLN representation of the relatively simple structures of inorganic compounds because too few keys are generated. Conversely, the more complex the molecule, the more involved the WLN, and inaccuracies or alternate representations may occur, possibly resulting in the erroneous generation of keys or failure to generate valid keys.

To demonstrate the use of WLN in molecular description, examples of assignment of WLNs to compounds are presented in Appendix B for an acyclic compound and in Appendix C for a cyclic compound. The need for accuracy in the generation of the WLN warrants emphasis, since WLN notation is the major factor in the equations for all four algorithms.

The next step in the translation procedure is to generate chemical descriptor keys for a compound based on the assigned WLN. This is accomplished by submitting the WLN to a computer program, developed by Enslein et al, as a part of this project, called Genkey 1/ Genkey 2.

Chemical descriptor keys provide an expression of molecular structure in terms of substructural fragments and lead to the development of quantifiable (key coefficient) values for use in the algorithms. Obtained from several sources, a total of 309 descriptor keys (with an additional 50 keys assigned based on the presence of certain combinations of keys 1-309) were used in the development of the four algorithms. None of the models employ all 359 keys in describing molecular structure. A subset of keys is generated by using statistical procedures that are described later in this report. Essentially, keys are selected by determining their contribution to the toxicity endpoint in question. This is determined by the frequency of the occurrence of a key (representing a specific molecular substructure) in the compounds listed in the modeling data base. In effect, the greater the frequency of occurrence the greater the probability that the key contributes significantly to the toxicity of that endpoint. Statistical methods are then used to calculate a coefficient value for each key in a

selected subset of the 359 possible keys. It is this quantifiable (numerical) representation of a molecular substructure that is used in the modeling equations to generate estimates of toxicity.

The number of keys selected from the 359 possible used in each model are as follows:

Endpoint	# Keys Selected
LD_{50}	- 82
Carcinogenicity	- 78
Mutagenicity	- 57
Teratogenicity	- 61

A list of all 359 keys and a description of the structure each represents is provided in Appendix D. A list of the keys in each model, their descriptions, and their coefficient values are provided as that model is described in this report.

Unfortunately the key generation programs are not error free. The contractor was unable to "de-bug" these computer programs within the time and funds allocated for this project. Three types of potential key generation problems are known to occur:

1. Keys not generated when they should be.

2. Keys generated when they should not be.

3. Keys erroneously generated. (Keys 310-350 represent certain combinations of keys 1-309 as defined in the description of each key presented in Appendix D). This is a particular problem with keys 311, 337, 342, and 349.

As a consequence, key files must be manually reviewed and compared against the WLN files for specific compounds to insure that all of the keys generated are correct on the basis of the assigned WLN. Corrections are made if necessary, and the data is resubmitted to the estimating program. Potential problem keys are listed in Table 2.

C. Statistical Methodologies

1. Selection of Variables.

For each modeling data base, variables to be included in defining the algorithm were determined using regression techniques (16). Stepwise regression or stepwise discriminant analysis as used based on whether the endpoint of the algorithm was considered as continuous or discriminant (3). If the endpoint was continuous, stepwise regression analysis was used. If the endpoint was discriminant (teratogen algorithm) discriminant analysis was used. As discussed later, it was necessary to use discriminate instead of regression analysis in developing the teratogen model because of the scoring process

TABLE 2. POTENTIAL PROBLEM KEYS

Problem Type	Key No.	Key Description	WLN Symbol(s)
1.	2	Positive charge	
	150	Chain primary amide	ZV or VZ
	151	Chain secondary amide	VM or MV
	152	Chain tertiary amide	N_V or VN
	181	Substituent primary amide	ZV or VZ
	182	Substituent secondary amide	VM or MV
	183	Substituent tertiary amide	N_V or VN
	162/193	Sulfonamide	(N)-SW or SW(N)
	163	Chain Guanidine	(N)-Y-U(N) or (N)-Y-U(N)-(N) or (N)UY-(N)-(N)
	165/196	Thioamide	SUYZ or YZUS
	166/197/304	Dialkylamino	(N)
	167/198	Methoxy	O1 or 1O
	171	Chain Phenylethyl	2R or R-(*)2
	172/203	Phenoxy	OR or R-(*)O
	178/209	Urea	(N)-V(N) Note: (N) can be in ring
	180	Biphenyl	R-(*)R
	189	Lactam	(N)V or V(N) within ring
	269	Potassium	-KA-
	306/309	Carbamate	OV(N) or (N)-VO
2.	154	Chain N-substituted acylhydrazide	Does not apply
	162	Chain sulfonamide	Does not apply
	163	Chain guanidine	Does not apply
	166	Chain dialkylamine (bonded to carbon)	Does not apply
3.	310-350	Refer to Appendix D for description	Does not apply

Note: (*) represents any locant; (N) represents any nitrogen.

From Enslein et al, (3).

used in determining teratogenicity of compounds in the modeling data base.

Stepwise regression procedures used to select variables may not always produce the best set of variables. The variables selected may be correlated and, as a result, produce a biased model (3, 17). To avoid such bias, candidate variables were selected from a larger set of variables, similar to those listed in Table 1, all of which were thought to make a possible contribution to the explanation of the statistical variance of the modeling data base (3). Ridge regression and a second stepwise regression were done using the candidate variables following the preliminary regression analysis.

The initial regression used a backward elimination procedure. All variables were included in the model and were selected out if their F-values were not significant at P=.05 (3, 17). In effect, candidate variables with low criterion or where F-values contributed least to the variance analysis equation were removed from the putative equation until the F-value reached was 1.7 (18). Ridge regression was performed on the remaining variables and ridge traces for each variable were examined to see whether any singularities existed which might suggest that the variable be omitted from the algorithm (3, 19). Least square estimates used in the backward elimination procedure might give results far removed from true variable values if the variables are correlated (17). The ridge regression was used to check the results of the stepwise regression. Finally, stepdown regression was repeated using only those variables retained following the ridge regression analysis.

In performing the regressions, outlier compounds (i.e., compounds that are not statistically characteristic of the main group of compounds) were identified and removed. The effect of removing a few outlier compounds from a large data set of several hundred compounds was felt to be minimal (3).

2. Statistical Evaluations of the Algorithms

Several statistical tests were used to evaluate the accuracy of classification by the algorithms. Of the evaluation tests used, the subset verification test was used to evaluate the accuracy of classification. This test probably provides the only practical evaluation of performance testing currently available (3). Using this test, a randomly selected subset of compounds is withheld from the data base that was created for the purpose of modeling the algorithm. The algorithm is then designed on the remaining compounds in the data base and is then tested with the subset of compounds set aside for that purpose. Residual plots, misclassification rates, and the Kilmogorov-Smirnov two-sample tests (18) were also used to test the model by comparing estimated values for endpoints with those values assigned based on actual values (i.e., reported bioassay testing) for the compounds in the verification test subset.

The results of the various statistical evaluation methods are presented following the description of the respective models. Statistical references cited should be consulted for a more detailed description of the statistical methods mentioned in this report.

D. Development of Individual Estimation Algorithms

1. General Modeling Considerations.

In developing all four algorithms, calculated data was easier to use if converted to equivalent logorithm values. Such conversion produces a normally distributed data base, (i.e., log-linear) and also eliminates the problem of dealing with a wide range of values such as 1:1000 which might occur in the dose ranges of 1 milligram to 1 gram seen in the LD_{50} algorithm. In the LD_{50} algorithm, the use of the reciprocal (1/C) of the reported or estimated LD_{50} concentration value creates a normal distribution of the data to facilitate the use of the logorithms. Consequently, to obtain the final estimated LD_{50} values in mg/kg or probability values between 0 and 1 for the other endpoints, it is necessary to take reciprocal values and convert back from logorithmic to actual values.

There are several steps (equations) necessary to obtain an LD_{50} estimate or probability value. Each algorithm, shown in its respective table, lists all of the descriptor keys (and, in the case of the LD_{50} and carcinogen algorithms, molecular weight) that have been found to be statistically significant to their toxic endpoints. The compound for which predictive toxicity estimates are to be generated is translated into the equivalent WLN. From the WLN, all keys that are represented in the WLN are selected from the total set of 359 keys. However, only those keys that also appear in the model subset of keys are used in calculating the positive and negative scores (e.g., carcinogen and noncarcinogen scores) for the carcinogen, mutagen, and teratogen algorithms or to calculate the estimated log (1/C) value in the LD_{50} algorithm (see Figure 3). These values are then used in the final estimating equation for each endpoint.

These equations are presented in a step-wise manner for each algorithm as it is discussed. An example use of each algorithm is presented in the appendices as indicated in the discussion of the models.

The LD_{50} algorithm expresses the estimated endpoint value as the dose of a compound, in mg/kg, necessary to kill one-half of the test animal population (i.e., lethal dose for 50% mortality, hence LD_{50}). The other three models express a predicted endpoint value within a range of 0.000 to 1.000 with 1.000, being the highest probability of the toxicologic endpoint occurring as a result of exposure to that compound (e.g., 0.989 probability of the compound being carcinogenic). For the purpose of this report, the terms probability and potential are

FIGURE 3. LD_{50} ESTIMATING EQUATION

The pertinent coefficient values (c) for each of the keys are summed (c) and added to the regression constant (0.552) and to the product of 0.681 x $\log_{10}$ (Mol. Wt.). The resulting value is the estimated log 1/c, where c is the number of moles of the compound which represents the LD_{50}.

$$\log (1/c) = .552 + .681 (\log_{10} \text{M.Wt.}) + c$$

To convert log 1/c to the estimated LD_{50}, expressed as mg/kg, use the following equation:

$$LD_{50} \text{ (mg/kg)} = \frac{1000 \times \text{M. Wt.}}{\text{antilog log (1/c)}}$$

considered interchangeable. The final equations of the algorithms developed are unusual in that they are expressed in tabular form because they are quite long and are not in the usually perceived algebraic form.

2. LD_{50} (oral, rat) Algorithm.

Data used in the LD_{50} algorithms originated from The Toxic Substances List (20) which is now called the NIOSH Registry of Toxic Effects of Chemical Substances (RTECS). The results of the LD_{50} algorithm are derived from a continuous (as opposed to a discriminant) endpoint, and the procedures for generating an estimated LD_{50} value are different from those of the other three models. These procedures are illustrated in Appendix E using an example compound.

There were two LD_{50} models developed in this project. An earlier model was based on 475 compounds selected from the letters A through M of the 1974 Toxic Substances List and 148 molecular substructure keys then available from the CROSSBOW program (3). The statistics for the equation of this algorithm are as follows:

Multiple correlation coefficient, R^2	.457
Standard error of estimated log (log 1/C + 1)	.089
Mean log 1/C	2.35
Standard deviation of log (log 1/C + 1)	0.68

With this equation, it was possible to predict the LD_{50} (oral, rat) of an untested compound so that approximately 63% of the compounds could be estimated within a factor of approximately 2.5, and virtually all compounds within a factor of 10 (in mg/kg units) (3). This, for example, means that an estimated oral rat LD_{50} dose of 1 mg/kg (with a factor of 2.5), when checked against actual reported data will correspond to a dose in the 0.25 mg/kg to 2.5 mg/kg range approximately 63% of the time.

In the second algorithm, 3,600 compounds were collected from the RTECS. This was essentially the entire population of compounds with oral rat LD_{50} data. This second algorithm was used to determine how many compounds would be needed in order to achieve stability of the structure-activity equation. Separate regression models were developed for three subsets of compounds of 1,000, 1,500, and 2,000 compounds as shown in Table 3. It was determined that there was very little change in the statistics associated with the model subsets of 1,500 and 2,000 compounds (3). Enslein et al, assumed that the major difference between these two models is due to the difference in the number of variables considered in these two models (77 for the earlier model and 103 for the later model) (3).

These results suggest "that at least for the available data, 2,000 compounds result in an essentially asymptotic equation" (3), (i.e., adding more compounds to the data set would not increase the strength of the equation).

TABLE 3. LD_{50} ALGORITHM: REGRESSION STATISTICS FOR SUBSET MODELS OF 1,000, 1,500, AND 2,000 COMPOUNDS

log 1/C

N	x	S.D.	S.E.	Skew	Kurtosis	Range	Residual Mean Square	P.F.	R^2	S.E. of Estimate
1,000	2.540	.860	.0272	.72	.72	.45-5.95	.36	892	.56	.60
1,500	2.540	.875	.0226	.71	.51	.45-5.90	.38	1,396	.52	.62
2,000	2.530	.880	.0197	.69	.52	.34-5.99	.39	1,864	.52	.62

From Enslein et al, (3)

The 2,000 compound model was therefore used as the basis for refining statistical procedures in the oral rat LD_{50} model (3). A complete list of these compounds is provided in Appendix F.

As a number of variables were removed from the equation, ridge regression analysis was performed. As shown in Table 4, residual plots (log 1/C actual - log 1/C predicted) produced from this regression analyses are poorly fitted at both the top and bottom ends (3). Note that the number of compounds dropped to 1,968 as a result of removing those found to be duplicates.

Because the range of residual plots values were poorly fitted in their distribution, it was necessary to compromise between range and fit in establishing a range of values with which to work (3). The range of values was limited to encompass log 1/C values between 1.25 and 4.75 in the final LD_{50} algorithm. This is considerably narrower than that in the first algorithm, which encompassed log 1/C values of approximately 1.0 to 6.2.

The LD_{50} algorithm presented in Table 5 includes all of the variables and their respective coefficient values as calculated by the statistical procedures described previously. The resulting equation for generating LD_{50} values based on this model is as shown in Figure 3.

A subset of 600 compounds were withheld for performance testing of the algorithm. Of these 600 compounds, 8 could not be properly processed by the WLN key generation program and 24 compounds were assigned none of the 82 keys present in the LD_{50} algorithm, leaving a test subset of 568 compounds. Log 1/C data for these 568 compounds were evaluated based on the equation presented in Table 5. Using a plot of the residual values (log 1/C actual - log 1/C predicted) as a function of the predicted values it was found that the prediction inaccuracies were greatest at the extremes of the range. This was not an unexpected finding, and because of this the predicted values were tabulated into ranges and statistics calculated for the compounds within each range. The results, presented in Table 6, show that there are no meaningful statistics available below log 1/C of 1.5 or above 4.0 (perhaps 3.5) (3). The standard deviation of the residuals from predicted log 1/C values from 1.5 to 3.5 varies between .58 and .81.

In examining the quantiles shown in Table 6, it is found that below mid-range there is a larger residual error for low values and above mid-range for the higher values. An example of the accuracy of the resulting estimates in the range of log 1/C of 2 to 2.5 is that 50% of the values between the semi-quartile range 25-75% would have an error of -.45 and +.34. As these are log values, they translate into actual LD_{50} values (i.e., mg/kg) lying between .355 and 2.19 times the estimated value. Similarly, 90% of the values are found between the 5th and 95th quantiles with an error range of -.87 to +1.03, which translates to the equivalent of .135 and 10.72 times the estimated values.

TABLE 4. DISTRIBUTION OF LOG 1/C FOR 1,968 COMPOUND ALGORITHM

log 1/C range	N	%
0.25 - 0.50	9	0.46
0.75 - 1.25	67	3.4
1.25 - 1.75	295	15.0
1.75 - 2.25	448	22.8
2.25 - 2.75	462	23.5
2.75 - 3.25	307	15.6
3.25 - 3.75	200	10.2
3.75 - 4.25	98	5.0
4.25 - 4.75	41	2.1
4.75 - 5 25	30	1.5
5.25 - 5.75	9	0.46
5.75 - 6.25	2	0.1

From Enslein et al, (3)

TABLE 5. LD_{50} ALGORITHM EQUATION

KEY	FREQUENCY	DESCRIPTION	COEFFICIENT	F
NON-CYCLIC PARTS OF MOLECULE				
K5	158	Terminal oxygen (not carbonyl)	.458	62.5
6	386	One 3-branch carbon atom	.096	6.8
9	21	Greater than 3-branch nitrogen atom.	.196	2.1
10	185	1 sulphur atom	.362	52.2
11	114	More than 1 sulphur atom	.821	150.7
14	168	1 double bond, excluding -C=S, -N=, or -C=O	.141	8.2
16	24	Triple bond	.189	2.4
CHAIN FRAGMENTS				
K17	338	1 methyl/methylene group	.089	5.8
20	223	Alkyl chain $(CH_2)_n$ or $CH_3(CH_2)_{n-1}$ where n=3-9	-.250	31.6
25	19	Bromine	.256	3.2
26	30	Fluorine	.435	11.4
28	153	One -NH- group	.334	29.1
30	102	One $-NH_2$ group	.236	14.5
31	20	More than one $-NH_2$ group	.258	3.5
34	54	Unusual carbon atom	.278	11.5
36	205	More than one -O- group	.211	14.4
37	195	One -OH group	-.163	12.2
43	123	One $-\overset{\overset{O}{\|}}{C}-O$ (ester) group	-.156	7.5
44	54	More than one $-\overset{\overset{O}{\|}}{C}-O$ (ester) group	-.205	5.7

TABLE 5. LD_{50} ALGORITHM EQUATION (Cont.)

KEY	FREQUENCY	DESCRIPTION	COEFFICIENT	F
SUBSTITUENT FRAGMENTS				
K47	90	Ethyl/ethylene group	.122	2.6
50	280	Generic halogen	.212	12.6
51	145	One chlorine	-.098	1.8
54	13	Fluorine	-.257	2.3
58	91	One $-NH_2$ group	.129	4.0
59	17	More than one $-NH_2$ group	.283	3.8
60	25	One -N= or HN= group	.270	5.1
66	24	More than one -OH group	.187	2.4
RING HETEROATOMS				
K75	20	Single occurrence of oxygen in more than one ring	.461	7.1
78	150	Multiple occurrence of nitrogen	.086	2.3
81	74	Single occurrence of sulphur	.140	3.7
82	7	Multiple occurrence of sulphur	.525	5.4
85	82	Single occurrence of carbonyl	-.122	2.5
90	4	Multiple occurrence of exocyclic double bond	.479	2.6
RING TYPES				
K99	100	Carbocyclic 6-membered ring	-.257	7.9
100	27	Carbocyclic ring other than 5 and 6-membered	-.198	2.2
104	233	1 heteroatom in one ring	.202	18.6
107	56	1 heteroatom in more than one ring	.477	15.4

TABLE 5. LD_{50} ALGORITHM EQUATION (Cont.)

KEY	FREQUENCY	DESCRIPTION	COEFFICIENT	F
RING FUSIONS				
K111	26	More than 1 single heterocyclic ring	-.574	15.6
112	65	1 single carbocyclic ring	.321	9.0
113	9	More than 1 single carbocyclic ring	.409	3.7
114	63	1 carbo/carbo fusion	.143	3.1
115	24	More than 1 carbo/carbo fusion	.373	7.1
120	4	1 carbo/hetero fusion in more than 1 ring system	-.935	8.8
123	5	More than 1 hetero/hetero fusion	.743	6.4
RING LINKAGE				
K130	99	Bilinkage	.271	13.9
EXTENSIONS				
K149	294	Presence of suffix	.098	4.1
ADDITIONAL CHAIN FRAGMENTS				
K151	17	Chain secondary amide	-.380	5.9
154	4	Chain N-substituted acylhydrazides	-.649	4.8
156	4	Chain amidine	.637	4.7
161	34	Chain N-nitroso	.194	2.8
162	3	Chain sulfonamide	-.680	3.7
165	35	Chain thioamide	.304	8.3
166	106	Chain dialkylamino	.255	16.4
167	35	Chain methoxy	-.272	7.1
171	19	Chain phenethyl	-.368	5.9
174	1	Chain phenylureido	.980	2.6

TABLE 5. LD_{50} ALGORITHM EQUATION (Cont.)

KEY	FREQUENCY	DESCRIPTION	COEFFICIENT	F
ADDITIONAL SUBSTITUENT FRAGMENTS				
K180	13	Biphenyl	-.432	5.8
182	37	Substituent secondary amide	-.244	5.9
188	12	Barbiturate	.316	3.1
193	14	Substituent sulfonamide	-.561	12.0
194	3	Substituent quanidine	.626	3.5
196	18	Substituent thioamide	.272	3.7
197	22	Substituent dialkylamino	.283	4.7
201	4	Substituent N-nitro	-.934	9.0
203	10	Substituent phenoxy	-.323	3.0
309	61	Substituent carbamate	.548	38.5
ADDITIONAL METAL FRAGMENTS				
K246	1	Fe	-1.435	6.0
250	2	Pb	-1.635	16.0
256	9	Hg	1.114	29.3
269	6	Ka	-.345	2.0
282	13	St	-.484	8.6
284	46	Na	-.266	7.1
293	22	Sn	1.175	78.6

TABLE 5. LD_{50} ALGORITHM EQUATION (Cont.)

KEY	FREQUENCY	DESCRIPTION	COEFFICIENT	F
CARCINOGENESIS KEYS				
K312	4	Organohalogen mustards	.864	8.2
315	161	Haloalkane	.323	21.2
322	4	Aziridine	.934	10.1
327	5	5-membered ring anhydrides	-.466	3.2
330	16	Fused aromatic - unsaturated lactone	.539	10.1
341	79	Aromatic nitro	.399	31.3
343	24	α,β-dihaloalkane	-.208	2.4
344	83	Geminal-dihaloalkane	-.315	10.0
348	3	Fused polychlorinated alicyclic	.578	2.8
350	18	Hydrazo/hydrazine	.324	5.0
LOG MOLECULAR WEIGHT			.681	53.5
CONSTANT			.552	

From Enslein et al, (3)

TABLE 6. TEST COMPOUNDS - CHARACTERISTICS OF RESIDUALS*

Predicted log 1/C	N	Quantiles												
		1	5	10	25	75	90	95	99	X	Median	S.D.	Min.	Max.
1.0 - 1.5	3	-.12	-.12	-.12	-.12	1.74	1.74	1.74	1.74	.72	.55	.94	-.12	1.74
1.5 - 2.0	78	-1.37	-1.05	-.71	-.42	.26	.54	.79	1.00	-.059	-.12	.66	-1.37	1.00
2.0 - 2.5	224	-1.34	-.87	-.67	-.45	.34	.76	1.03	1.74	-.015	-.059	.58	-1.49	1.94
2.5 - 3.0	143	-1.99	-1.25	-.78	-.45	.36	.86	1.45	2.71	-.017	-.07	.77	-2.33	2.76
3.0 - 3.5	97	-1.80	-1.20	-.81	-.46	.52	1.17	1.56	2.51	.081	.067	.81	-1.80	2.51
3.5 - 4.0	18	-1.63	-1.63	-1.60	-1.07	.50	1.48	1.64	1.64	-.21	-.35	.98	-1.63	1.64
4.0 - 4.5	4	-1.02	-1.02	-1.02	-1.02	.85	.85	.85	.85	-.082	-.077	1.07	-1.02	.85
	567													

* Residual values = log 1/C actual - (log 1/C predicted)

From Enslein et al, (3)

It is difficult to know which fraction of the residuals in the model fit inadequately due to the model itself, or as a result of other factors such as inadequately measured LD_{50} values or discrepancies between data resulting from replication studies between different laboratories (3). Enslein et al, found that one of the compounds in RTECS incorrectly reported an LD_{50} value of 70 ug/kg instead of 70 mg/kg. This compound was dropped from the data shown in Table 6. Despite such limitations, it would seem that this model can generate LD_{50} estimate values at least as well as those reported thus far in the literature. However, there have been insufficient numbers of compounds for which extensive replications have been carried out to be able to make such a statement with a great deal of confidence (3).

3. Mutagenicity Algorithm

Compounds incorporated in the data base for developing this model were obtained by screening the files of the Environmental Mutagen Information Center (EMIC), Oak Ridge National Laboratory, Oak Ridge, Tennessee, and reports from the National Toxicology Program (NTP) for all compounds for which Ames test for mutagenicity data had been reported. Essentially, all publications from the EMIC files relating to the Ames Test for mutagenicity (encompassing over 1200 compounds) were reviewed and the test results recorded. Judgments as to the quality of the reported data, e.g., in terms of dose response reported, were made by contractor chemists and toxicologists. In general, the Gene-Tox Criteria (21) were applied in this subjective evaluation of the data. Using these criteria, a compound was classified as a nonmutagen if it had been tested with negative results in at least three of the five strains of Salmonella Typhimurium (TA98, TA100, TA1535, TA1537, and TA1538) used in the Ames Test. For a compound to be classified mutagenic it had to be tested with positive results in at least two of the five strains. It should be noted that two of the five strains (TA98 and TA100) are considered less sensitive than the other three in assessing mutagenicity and, therefore, weighed less in the decision to classify a compound as mutagenic (22).

The chemical selection committee of the National Toxicology Program (NTP) uses the Gene-Tox criteria but requires at least four strains instead of three to be tested and a negative result in all four strains for a compound to be classified as a nonmutagen. Additionally, NTP also requires that each of the tests be repeated in at least one other laboratory. In the case of compounds with conflicting data, decisions regarding positive or negative mutagenic classification were made only if test results among at least two different laboratories were mutually reinforcing. When conflicting results could not be so resolved, the compound was discarded from the data base and subsequent modeling and testing procedures. Because of the more stringent requirements, an NTP judgment was held to supercede those obtained from EMIC.

After applying these criteria to over a thousand compounds, a total of 301 were judged to be positive mutagens and 231 to be nonmutagens. From these two groups a subset of 37 positive and 23 negative compounds were randomly selected and set aside for subset verification testing. A list of all the compounds used in the mutagen modeling data base are presented in Appendix G.

The equations for the mutagen algorithm were derived by discriminant analysis and ridge regression procedures. Based on the mutagenicity model equation presented in Table 7, a mutagen and nonmutagen score is calculated using the equations shown in Figure 4. The regression constant values (of -5.078 and -3.183) result from the regression analysis applied to the mutagen and nonmutagen groups of compounds in calculating the coefficient values for the keys selected for the mutagen algorithm. The equation for generating an estimate of mutagenicity incorporates the resulting exponential values of these equations. Natural log values are determined for both score values and then used in the probability equation shown in Figure 5. A step-by-step illustration of this procedure is presented in Appendix H using an example compound.

Several methods exist to evaluate the accuracy of the mutagenesis model. The simplest method is to indicate the number of compounds correctly classified by the model. As shown in Table 8, varying the range of the indeterminate (i.e., cannot sufficiently discriminate between mutagen or nonmutagen) zone between p = .4 to .599 and p = .3 to .699 the percent of false positives and false negatives increases as the indeterminate range decreases. The wider the indeterminate range, the larger the number of compounds which cannot be classified, in addition to some reduction in the number of misclassified compounds.

A second method for estimating accuracy is to compare the actual error rate in specified probability ranges to the expected error rate (based on the binomial distribution). These data are shown in Table 9. Using a two sample Kolmogorov-Smirnov test (18), it was found that there was not a statistically distinguishable difference in the actual and expected cumulative error distributions (3). This would indicate that the probability values derived from this model have a high degree of precision and can be used with confidence for the ranking of compounds (3). The results of this statistical test were not made available to the author in the draft report provided by the contractors, and thus are not presented here.

The subset test provides a third way to assess the accuracy of this model. As described previously in the LD_{50} model, a number of compounds for which mutagenesis data had been obtained were held back from the modeling set by a random selection process. These compounds were then evaluated by means of the discriminant equation of the model and the probability values of mutagenicity were compared to the reported values. As seen in Table 10, the results of the test parallel those results shown in Table 8 with the exception that a larger number of compounds

TABLE 7. MUTAGEN ALGORITHM EQUATION

KEY	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	Positive	Negative		Positive	Negative	Difference Pos-Neg	F
NON-CYCLIC PARTS							
K3	60	12	Branching terminal nitro-group $-NO_2$	5.001	0.843	4.158	65.5
5	48	3	Terminal oxygen (not carbonyl)	4.946	1.819	3.126	14.3
8	55	14	3-branch nitrogen atom	1.668	-0.489	2.156	8.8
10	13	10	1 sulphur atom	2.498	0.621	1.877	4.9
11	5	3	More than 1 sulphur atom	4.032	0.549	3.484	5.7
14	14	13	1 double bond, excluding -C=S, -N=, or -C=U	3.746	1.209	2.537	11.1
CHAIN FRAGMENTS							
K18	37	42	More than 1 methyl/methylene group	1.215	2.124	-1.000	4.1
19	33	26	Ethyl/ethylene group	0.234	2.595	-2.361	15.3
20	8	14	Alkyl chain (CH_2) or CH_3 $(CH_2)_{n-1}$	-0.050	2.814	-2.864	12.0
26	0	5	Fluorine	-4.801	2.994	7.795	14.0
29	2	3	More than one -NH- group	3.091	-0.161	3.252	2.2
31	3	0	More than one $-NH_2$ group 9.230	2.561	6.669	9.9	
35	9	6	One -O- group	2.096	-1.151	3.247	10.1
37	26	7	One -OH group	2.792	1.602	1.191	2.6
38	9	6	More than one -OH group	3.335	1.873	1.462	1.9
41	1	18	One -C-OH (acid) group	0.757	5.136	-4.379	25.4
151	1	2	Chain secondary amide	-4.458	0.956	-5.415	6.1
156	1	0	Chain amidine	4.457	-4.243	8.700	5.1
161	22	0	Chain N-nitroso	0.973	-2.569	3.542	10.2

TABLE 7. MUTAGEN ALGORITHM EQUATION (Cont.)

KEY	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	POSITIVE	NEGATIVE		POSITIVE	NEGATIVE	DIFFERENCE POS-NEG	F
163	2	5	Chain guanidine	-0.123	4.729	-4.852	10.2
168	3	0	Chain hydroxylamine	7.410	-4.683	12.092	15.0
178	5	2	Chain urea	-1.043	2.104	-3.147	3.6
SUBSTITUENT FRAGMENTS							
K48	0	1	Alkyl chain $(CH_2)_n$ or $CH_3(CH_2)_{n-1}$ where n = 3-9	-2.749	4.418	-7.167	3.9
53	0	6	Bromine	-1.898	1.707	-3.605	5.6
54	9	0	Fluorine	4.830	1.051	3.779	9.2
55	0	2	Iodine	-2.470	3.944	-6.414	5.5
56	20	9	One -NH- group	0.581	-2.415	2.997	6.6
58	23	8	One $-NH_2$ group	2.186	-0.482	2.668	8.0
59	13	2	More than one $-NH_2$ group	5.379	-0.619	6.000	26.0
66	23	14	More than one -OH group	2.551	-0.018	2.569	12.4
69	4	18	One -C-OH (acid) group	1.314	3.287	-1.973	5.7
180	18	1	Biphenyl	4.747	-0.389	5.136	13.8
182	4	5	Substituent secondary amide	-4.585	2.709	-7.293	18.6
190	9	3	Substituent azo and diazo	1.677	-2.865	4.543	13.5
193	0	3	Substituent sulfonamide	-4.656	0.046	-4.702	3.9
195	2	0	Substituent N-N	4.222	-0.100	4.321	2.8
200	1	0	Substituent oxime	9.191	-0.301	9.491	6.5
209	2	5	Substituent ureas	-2.535	3.595	-6.130	13.5

TABLE 7. MUTAGEN ALGORITHM EQUATION (Cont.)

KEY	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	POSITIVE	NEGATIVE		POSITIVE	NEGATIVE	DIFFERENCE POS-NEG	F
RING HETEROATOMS							
K73	51	24	Single occurrence of oxygen	3.250	1.180	2.070	13.1
78	17	11	Multiple occurrrence of nitrogen	3.192	2.136	1.057	1.8
RING TYPES							
K99	20	26	Carbocyclic 6-membered ring	-1.976	1.971	-3.947	35.1
103	32	11	Heterocyclic rings other than 5 and 6-membered	4.139	0.743	3.396	24.7
HETEROATOM COUNT							
K106	2	3	More than 2 heteroatoms in one ring	-3.387	2.065	-5.452	9.2
RING FUSIONS							
K114	12	7	1 carbo/carbo fusion	1.231	2.792	-1.560	1.7
116	0	2	1 carbo/carbo fusion in more than 1 ring system	-7.864	4.431	-12.295	20.8
RING LINKAGE							
K127	2	1	True bridge indicator	5.589	1.617	3.973	3.3
UNUSUAL CONDITIONS							
K133	3	2	Inorganics	8.707	4.423	4.284	6.6

TABLE 7. MUTAGEN ALGORITHM EQUATION (Cont.)

KEY	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	POSITIVE	NEGATIVE		POSITIVE	NEGATIVE	DIFFERENCE POS-NEG	F
TOTAL RING FEATURES							
K137	50	68	1 benzene ring	2.134	3.771	-1.637	12.8
138	29	24	2 benzene rings	1.300	4.140	-2.839	17.7
139	0	1	More than 2 benzene rings	-7.093	2.664	-9.757	6.8
141	27	12	2 carbocyclic rings	3.410	1.100	2.313	7.5
ADDITONAL METAL FRAGMENTS							
K216	0	1	As	-5.409	1.605	-7.014	3.7
256	0	1	Ta	-4.909	2.446	-7.355	3.9
288	0	1	Te	-2.391	4.585	-6.976	3.4
CARCINOGENIC KEYS							
K310	49	17	Aromatic amino	1.820	0.474	1.346	2.4
315	23	8	Haloalkane	6.365	0.230	6.136	47.9
327	0	2	5-membered ring anhydrides	0.293	6.060	-5.767	5.1
331	38	9	Fused polynuclear aromatic	6.041	2.032	4.009	35.0
344	3	6	Geminal-dihaloalkane	-3.996	-0.432	-3.565	3.8
CONSTANT				-5.078	-3.183	-1.894	

From Enslein et al, (3)

FIGURE 4. EQUATIONS FOR CALCULATING MUTAGEN AND NONMUTAGEN SCORES

e^{expt+} = Mutagen Score = sum of coefficient values of assigned keys + regression constant.

e^{expt-} = Nonmutagen Score = sum of coefficient values of assigned keys + regression constant.

Natural log values are determined for both score values and then used in the probability estimating equation presented in Figure 5.

FIGURE 5. MUTAGENICITY ESTIMATING EQUATION

$$\text{Probability of Mutagen} = \frac{e^{expm+}}{e^{expm+} + e^{expm-}}$$

TABLE 8. MUTAGEN ALGORITHM - DESIGN COMPOUNDS

		Classification by discriminant equation					
		Positive		Indeterminate		Negative	
		N	%	N	%	N	%
Indeterminate: 0.400 - 0.599							
"ACTUAL" CLASSIFICATION	Positive	230	87.1	9	3.4	25	9.5
	Negative	22	10.6	11	5.3	175	84.1
Indeterminate: 0.300 - 0.699							
"ACTUAL" CLASSIFICATION	Positive	218	82.6	25	9.5	21	8.0
	Negative	13	6.3	27	13.0	168	80.8

From Enslein et al, (3)

TABLE 9. MUTAGEN ALGORITHM - MISCLASSIFICATION IN RANGES

Probability of Mutagenicity	No. of Compounds in range	Proportion Misclassified	N	Actual Cumulative	Expected Cumulative
.9 - 1.000	185	.027	5	5	9.25
.8 - .899	28	.071	2	7	13.45
.7 - .799	18	.033	6	13	17.95
.6 - .699	21	.429	9	22	25.3
.5 - .599	12	.500	6	28	30.7
.4 - .499	8	.375	3	31	34.3
.3 - .399	11	.364	4	35	38.15
.2 - .299	13	.538	7	42	41.4
.1 - .199	35	.343	12	54	46.65
.0 - .099	141	.014	2	56	53.7

From Enslein et al, (3)

TABLE 10. MUTAGEN ALGORITHM - TEST COMPOUNDS

		Classification by discriminant equation					
		Positive		Indeterminate		Negative	
		N	%	N	%	N	%
Indeterminate: 0.400 - 0.599							
"ACTUAL" CLASSIFICATION	Positive	26	81.2	1	3.1	5	15.6
	Negative	3	13.0	2	8.7	18	78.3
Indeterminate: 0.300 - 0.699							
"ACTUAL" CLASSIFICATION	Positive	25	78.1	5	15.6	2	6.3
	Negative	2	8.7	6	26.1	15	65.2

From Enslein et al, (3)

fall in the indeterminate range. In estimating the mutagenicity of these test compounds, it was observed that compounds which are essentially two-dimensional, (i.e., planar) can intercalate with DNA structures and are apparently not good subjects for the discriminant equation (3). These compounds were mostly found to be those with benzene rings and were removed from the test data set (3).

4. Carcinogenicity Algorithm

Compounds selected for the development of this model were obtained from volumes 1 through 17 of the International Agency for Research on Cancer (IARC) (23). However, not all of the compounds included in these volumes were used in the modeling data base for reasons which include; compounds with unidentified structure, polymers, compounds with unresolved WLN coding errors, compounds for which substructure keys could not be generated, and compounds "which were manifestly too different from the remainder to be included in the same data base" (3).

IARC uses a six category classification system for classifying the carcinogenicity of compounds (23). These six categories were:

1. Definite human carcinogen
2. Definite animal carcinogen
3. Suspect human carcinogen
4. Suspect animal carcinogen
5. Indefinite human carcinogen
6. Noncarcinogen

Sometime ago IARC eliminated the last category. Therefore, for the purposes of this project, a compound can be either a carcinogen or an indefinite carcinogen. Accordingly, in developing this model, categories 1 and 2 were combined as "definite carcinogens", categories 3 and 4 were discarded because they represented indeterminate data, and all other data were classified as "indeterminate carcinogens". The model was then developed on the basis of these two new categories of compounds. Of the total 406 listed in the IARC monographs 1-17, 223 compounds were listed as definite carcinogens and 120 listed as indefinite carcinogens. A complete list of these compounds is provided in Appendix I.

In this model, the selection of relevant descriptor keys from the total set of 359 keys involved initial withholding of molecular weight as a descriptive parameter. This was because molecular weight had been statistically determined to play a very significant role in the overall design of the carcinogen model (3). By holding this parameter back until the key selection process was completed, undue influence of molecular weight as a factor was avoided in the selection of the keys relevant to the model.

Based on the model equation shown in Table 11, the equations shown in Figure 6 are used to calculate a definite and indefinite score. These scores, as shown in Figure 7 are then used to generate the estimate of carcinogenic probability. A step-by-step illustration of this procedure is presented in Appendix J, using an example compound.

Several methods were used to evaluate the accuracy of this model. The simplest method is to indicate the number of compounds correctly identified by the model. As shown in Table 12, the percent of false positive and false negative probability estimates vary with a change in the range of the indeterminate zone. With an indeterminate zone ranging from p = .4 to .599 the false negative rate is 4.9% and the false positive rate is 11.7% with no decision available for 19 (5.5%) of the compounds. Increasing the indeterminate zone to a range of .3 to .699 produces a false negative rate of 10.8% and a false positive rate of 3.6% with no decision available for 35 (10.2%) of the compounds.

Another method used in evaluating model estimates was to compare the actual error rates to the expected error rates (based on a binomial distribution) within specified ranges. These data are shown in Table 13.

Using a two-sample Kolmogorov-Smirnov test (18) provides a third method for evaluating the model. This test found no statistically distinguishable difference between the actual and expected cumulative error distributions, leading to the conclusion that the probabilities derived from this model have a high degree of precision and can be used with confidence for the ranking of compounds (3). The data for this evaluation were not made available to the author in the contract draft report and are therefore not included in this report.

5. Teratogen Algorithm

Data for this algorithm were obtained from several sources. These sources included the texts, Catalog of Teratogens (24) and Drugs as Teratogens, (25) as well as the files of the Environmental Teratogen Information Center (ETIC) at the Oak Ridge National Laboratory at Oak Ridge, Tennessee. The criteria for evaluating and selecting compounds from these sources were as follows:

a. A compound must have test data available on at least two different mammalian species.

b. Test data must have been derived and reported by knowledgeable and competent sources.

c. Test data must have been obtained after 1969.

The second criterion of knowledgeable and competent source was applied subjectively by contractor chemists and toxicologists

TABLE 11. CARCINOGEN ALGORITHM EQUATION

Key	Description	Coefficients for Group		Difference	
		Indefinite	Definite	Definite - Indefinite	F
NON-CYCLIC PARTS					
5	Terminal oxygen (not carbonyl)	3.04	-0.16	-3.20	7.40
8	3-branch nitrogen atom	-0.63	4.19	4.82	20.5
11	> 1 sulphur atom	-8.53	-4.17	4.36	3.18
13	> 1 -C=S group	20.88	3.00	-17.88	20.8
CHAIN FRAGMENTS					
17	1 methyl/methylene group	3.66	1.98	-1.68	4.25
18	> 1 methyl/methylene group	3.46	0.80	-2.66	11.3
20	Alkyl chain $(CH_2)_n$ or CH_3 $(CH_2)_{n-11}$, n = 3-9	1.63	-0.39	-2.02	2.18
22	Generic halogen	3.16	4.98	1.82	2.65
23	1 chlorine	-0.28	2.97	3.25	7.39
28	1 -NH- group	0.24	-2.71	-2.95	2.71
30	1 $-NH_2$ group	3.97	6.41	2.44	4.45
35	1 - O- group	4.57	7.74	3.17	4.90
36	> 1 - O- group	2.45	6.10	3.65	7.39
38	> 1 - OH group	1.79	-1.15	-2.94	5.45
39	1 -C=O group	-1.54	0.37	1.91	2.20
ADDITIONAL CHAIN FRAGMENTS					
152	Tertiary amide	2.66	-3.83	-6.49	7.34
161	N-nitroso	1.80	5.44	3.64	6.07
164	N-N, azoxy	5.34	8.61	3.27	2.21
167	Methoxy	-4.05	-0.46	3.60	3.63
172	Phenoxy	4.82	-3.59	-8.41	11.2
173	Phenylazo and phenylhydrazo	0.85	-7.00	-7.85	3.37

TABLE 11. CARCINOGEN ALGORITHM EQUATION (Cont.)

Key	Description	Coefficients for Group: Indefinite	Coefficients for Group: Definite	Difference: Definite - Indefinite	F
SUBSTITUENT FRAGMENTS					
51	1 Chlorine	4.24	-1.17	-5.41	10.7
60	1 -N= or HN=group	-6.52	3.27	9.79	16.7
61	> 1 -N= or HN=group	-3.06	7.35	10.41	12.3
65	1 -OH group	2.55	0.53	-2.02	4.61
69	-C-OH (acid) group	1.46	-3.85	-5.31	7.71
71	-C-O (ester) group	0.18	-4.53	-4.71	6.71
ADDITIONAL SUBSTITUENT FRAGMENTS					
188	Barbiturate	-10.07	0.35	10.42	8.91
190	Azo and diazo	4.41	-6.60	-11.01	16.1
198	Methoxy	-2.04	0.48	2.52	2.47
204	Phenylazo and phenylhydrazono	6.92	-4.54	-11.46	18.0
RING HETEROATOMS					
75	Single occurrence of oxygen in 1 ring	9.12	0.80	-8.32	13.1
78	Multiple occurrence of nitrogen in 1 ring	4.87	0.96	-3.91	5.02
79	1 occurrence of nitrogen in 1 ring	7.71	-3.35	-11.06	17.4
86	Single occurrence of carbonyl	1.90	-5.43	-7.33	21.4
89	Single occurrence of exocyclic bond	-5.83	-1.38	4.45	6.00
90	> 1 exocyclic bond	3.03	-5.78	-8.81	11.7
94	> 1 occurrence of any letter other than (H,K,M,N,O,S,T,U,V,X,Y)	-0.45	-9.37	-8.92	3.09

TABLE 11. CARCINOGEN ALGORITHM EQUATION (Cont.)

Key	Description	Coefficients for Group: Indefinite	Coefficients for Group: Definite	Difference: Definite - Indefinite	F
UNUSUAL CONDITIONS					
131	Chelate	-3.23	3.68	6.91	2.38
TOTAL RING FEATURES					
134	1 ring system	7.18	3.55	-3.63	10.9
137	1 benezene ring	5.83	1.32	-4.51	38.1
143	1 heterocyclic ring	0.96	4.89	3.93	4.38
METAL FRAGMENTS					
216	As	8.65	0.33	-8.32	16.0
227	Cr	8.28	0.13	-8.15	17.1
246	Fe	10.60	-2.52	-13.12	19.8
250	Pb	3.79	-0.59	-4.38	6.54
261	Ni	5.89	9.18	3.29	2.52
282	Si	12.39	1.16	-11.23	10.2
284	Na	3.62	-1.98	-5.60	10.6
285	Sr	-10.34	3.59	13.93	8.43
301	Zn	-7.06	2.63	9.69	8.03
FDA SUSPECTED STRUCTURES					
310	Aromatic amino	4.42	1.84	-2.58	10.6
312	Organohalogen mustard	-2.82	-8.28	-11.10	6.62
314	Organohalogen mustard	-17.13	-6.03	11.10	5.00
318	Halogenated aromatic	0.17	-5.83	-6.00	9.76
321	Epoxide	3.86	-4.15	-8.01	17.5
324	β -lactone	-2.02	5.47	7.49	5.71
326	β -unsaturated lactone	10.17	4.65	-5.52	2.90
329	γ - β unsaturated lactone	-3.94	0.68	4.62	2.35
331	Fused polynuclear aromatic	3.66	-0.08	-3.74	9.22

TABLE 11. CARCINOGEN ALGORITHM EQUATION (Cont.)

Key	Description	Coefficients for Group: Indefinite	Coefficients for Group: Definite	Difference: Definite - Indefinite	F
RING TYPES					
98	Carbocyclic 5-membered ring	1.07	-2.39	-3.46	3.89
99	Carbocyclic 6-membered ring	-2.39	-0.59	1.80	2.19
101	Heterocyclic 5-membered ring	-4.10	1.45	5.55	12.1
102	Heterocyclic 6-membered ring	-4.33	2.18	6.51	12.7
103	Heterocyclic rings other than 5 and 6-membered	-4.42	7.20	11.62	30.3
HETEROATOM COUNT					
104	1 heteroatom in 1 ring	0.79	-3.73	-4.52	11.8
RING FUSIONS					
110	1 single heterocyclic ring	1.72	-1.60	-3.32	4.58
113	> 1 single carbocyclic ring	7.67	-8.57	-16.24	17.7
114	1 carbo/carbo fusion	-10.65	1.05	11.70	41.0
115	> 1 carbo/carbo fusion	-2.58	3.94	6.52	16.4
119	> 1 carbo/hetero fusion	-0.03	-6.67	-6.64	14.6
122	1 hetero/hetero fusion	-13.12	2.52	15.64	31.7
RING LINKAGES					
127	True bridge indicator	11.15	-2.95	-14.10	13.8
128	1 multi-cyclic point	-1.57	3.00	4.57	4.94
129	> 1 multi-cyclic point	-3.17	-0.22	2.95	2.19
130	Bilinkage	2.44	4.12	1.68	3.22

TABLE 11. CARCINOGEN ALGORITHM EQUATION (Cont.)

Key	Description	Coefficients for Group: Indefinite	Coefficients for Group: Definite	Difference: Definite - Indefinite	F
OTHER COMBINATION KEYS					
335	One or more occurrences of keys 100,207,223,285,314,330,332	-6.93	2.17	9.10	16.9
336	One or more occurrences of keys 21,42,94,176,281,309	6.37	0.75	-5.62	10.6
MOLECULAR WEIGHT		0.01062	0.02046	.00984	14.6
CONSTANT		-7.42	-6.55	0.87	

From Enslein et al, (3).

FIGURE 6. EQUATIONS FOR CALCULATING DEFINITE (CARCINOGEN) AND INDEFINITE (NONCARCINOGEN) SCORES

e^{expt+} = Carcinogen score (definite) = sum of coefficient values of assigned keys + regression constant.

e^{expt-} = Noncarcinogen score (indefinite) = sum of coefficient values of assigned keys + regression constant.

Natural log values are determined for both score values and then used in the probability estimating equation presented in Figure 7.

FIGURE 7. CARCINOGENICITY ESTIMATING EQUATION

$$\text{Probability of Carcinogen} = \frac{e^{expt+}}{e^{expt+} + e^{expt-}}$$

TABLE 12. CARCINOGEN ALGORITHM - CLASSIFICATION BY DISCRIMINANT EQUATION

		Classification by Discriminant Equation					
		Non or indefinite		Indeterminate		Definite	
	Indeterminate: .4 - .599	N	%	N	%	N	%
IARC Carcinogen Classification	Non or Indefinite	96	80.0	10	8.33	24	11.7
	Definite	11	4.9	9	4.0	203	91.0
	Indeterminate: .3 - .699						
IARC Carcinogen Classification	Non or Indefinite	93	77.5	14	11.7	13	10.8
	Definite	8	3.6	21	9.4	194	87.0

From Enslein et al, (3)

TABLE 13. CARCINOGEN ALGORITHM - MISCLASSIFICATION IN RANGES

Probability of Carcinogenicity	No. of Compounds in range	Proportion Misclassified	Actual N	Actual Cumulative	Expected Cumulative
.9 - 1.0	171	.0292	5	5	8.55
.8 - .899	25	.24	6	11	12.3
.7 - .799	11	.182	2	13	15.05
.6 - .699	10	.10	1	14	18.55
.5 - .599	11	.545	6	20	23.5
.4 - .499	8	.5	4	24	27.1
.3 - .399	6	.5	3	27	29.2
.2 - .299	7	.286	2	29	30.95
.1 - .199	11	.273	3	32	32.60
.0 - .099	83	.0361	3	35	36.75

From Enslein et al, (3)

based on their own opinions and expertise. The date criterion was applied because the methods of testing, evaluating, and reporting on teratogenic data had become relatively standardized at this time, and more uniform consideration and evaluation of the reported data was possible (3). These criteria were obtained from and discussed with Dr. Bryan Hardin of NIOSH, Cincinnati, and served to provide some quality control of the data used in creating the modeling data base.

In creating the teratogen modeling data base, each compound selected was scored on a scale of 0.0 to 1.0, zero meaning no evidence of teratogenicity and 1.0 meaning definite evidence of human teratogenicity (e.g., thalidomide). This evaluation and classification procedure was used for the teratogen algorithm because there was no classification process currently available for categorizing compounds with this toxic endpoint. Table 14 provides a more detailed definition of the scoring procedure used in this screening process. The scoring was accomplished by having the list of compounds under consideration evaluated by four teratologists or toxicologists, allowing for additional comments regarding the validity of scores. These reviewers were:

Dr. Bryan Hardin, NIOSH, Cincinnati, Ohio.
Dr. Jeanne Manson, University of Cincinnati, Cincinnati, Ohio.
Dr. Orville Paynter, U.S. Department of Commerce,
Washington, D.C.
Dr. James Schardein, Warner-Lambert/Parke-Davis,
Ann Arbor, Michigan.

Differences of opinion in this scoring process were resolved or the compound was dropped from consideration.

Approximately 670 compounds were selected from the data sources listed. In the discriminant analysis, two groups of compounds were used. On an arbitrary basis, those with scores between 0 and .25 were labeled nonteratogens, and those with scores between .75 and 1.0 were labeled teratogens. As a result of this grouping of compounds, the final modeling data base consisted of 235 teratogens and 191 nonteratogens for a total of 426 compounds. The complete list of compounds is presented in Appendix K.

The equations in this algorithm are similar to those of the mutagen and carcinogen algorithms. As shown in Figure 8, indefinite and definite scores are also calculated for each compound, used in the estimation of teratogenic probability as shown in Figure 9. A step-by-step illustration of this procedure is presented in Appendix L, using an example compound.

Unlike the mutagen and carcinogen models, attempts at developing a teratogen model using regression methods were not successful (3). This was attributed to the number of assigned scores near 0.5 in the scoring process which might affect least-square estimates produced by linear regression. As a result, efforts were concentrated on a discriminant analysis approach for the teratogenesis model.

TABLE 14. CRITERIA FOR EVALUATION OF TERATOGENICITY*

*	0.0	most probably not teratogenic, negative in two or more species.
*	.01 to .20	no evidence, or positive in obscure species, unconfirmed.
*	.21 to .40	positive in one species, questionable in second species - data may be suspect.
*	.41 to .60	equal or near equal evidence pro and con, good data in one species only.
*	.61 to .75	teratogenic in two species - studies could be better.
*	.76 to .85	teratogenic in two species or in monkey or in good case study.
*	.86 to .99	teratogenic in two species - fairly good evidence in human.
*	1.0	no doubt of its teratogenicity in humans.

* These evaluation criteria were developed and applied by contractor chemists and toxicologists as well as those acknowledged in the text of this report; the resulting evaluations reflect their subjective opinions and expertise and not necessarily those of the author of this report.

FIGURE. 8 EQUATIONS FOR CALCULATING TERATOGEN AND NONTERATOGEN SCORES

e^{expt+} = Teratogen Score = sum of coefficient values of assigned keys + regression constant

e^{expt-} = Non-Teratogen Score = sum of coefficient values of assigned keys + regression constant

Natural log values are determined for both score values and then used in the probability estimating equation presented in Figure 9.

FIGURE 9. TERATOGENICITY ESTIMATING EQUATION

$$\text{Probability of Teratogen} = \frac{e^{expt+}}{e^{expt+} + e^{expt-}}$$

In the discriminant analysis model two groups of compounds were used. One group, with scores between 0.00 and 0.25 were considered as being non-teratogens, and a second group, with scores between 0.75 and 1.00 as being teratogenic. Following the identification of outlier compounds, misidentified compounds and the application of statistical procedures as previously described the model equation as shown in Table 15 was developed. This model is based on 195 compounds in the non-teratogenic group and 235 compounds in the teratogenic group. A list of all compounds in the modeling data base and their scores are listed in Appendix L. The teratogen model is as presented in Table 16.

As with the other models, several methods were used to evaluate the accuracy of classification of this model. As illustrated in Table 17 false positive and false negative rates vary with changes in the range of the indeterminate zone. At an indeterminate range of .40 to .599 the false positive rate is approximately 14% and the false negative rate is approximately 13% with approximately 8% of the compounds considered indeterminate. With a wider indeterminate range of .30 to .699 the approximate rate for false positives is 8% versus a false negative rate of 10% with 22% of the compounds classified as indeterminate. Note that Table 17 reflects only 426 of the 430 compounds originally in the modeling data base. This is because after the equations had been developed, four compounds were found to have been misscored and were removed from the data base.

Table 18 evaluates the accuracy of this algorithm by presenting the misclassifications in various probability ranges. Note that in the extreme ranges the accuracy of classification is very high. The misclassification rate, for example, in the probability range of 0.9 to 1.0 is less than 1%.

A two-sample Kolmogorov-Smirnov test (18) was applied to the data presented in Table 18 resulting in a finding that the misclassification distribution was not statistically distinguishable from the expected distribution. Based on this result, it was concluded that no serious bias existed in these equations (3). Again, the results of this evaluation were not made available by the contractor and cannot be presented in this report.

III. Estimation and Ranking of NOHS Compounds

Predicted estimates for one or more of the four toxicologic endpoints were generated for a number of compounds listed in the NOHS data base. Estimates were not generated for all compounds listed for one or more of the following reasons:

--- Compound already in the respective model.

--- Toxicity data for compound already exists in RTECS.

--- A WLN could not be generated for compound.

TABLE 15. TERATOGEN ALGORITHM EQUATION

KEY	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	NON-TERATOGENS	TERATOGENS		NON-TERATOGENS	TERATOGENS	DIFFERENCE TER.-NONTER.	F VALUE
ALL PARTS OF MOLECULE							
K2	7	0	Positive charge	3.334	-0.995	-4.329	12.4
ALL NON-CYCLIC PARTS OF MOLECULE							
K5	3	12	Terminal oxygen	0.445	3.533	3.088	10.7
7	15	3	4-branch carbon atom	2.491	-0.386	-2.877	11.1
10	20	22	1 sulphur atom	-0.458	1.826	2.284	11.7
12	1	5	1 -C=S group	0.180	4.272	4.092	8.9
15	1	4	More than 1 double bond, excluding -C=S, -N=, or -C=O	-4.292	1.709	6.001	16.0
CHAIN FRAGMENT							
K18	62	51	More than 1 methyl/ Methylene group	4.016	2.306	-1.710	17.1
19	34	36	Ethyl/ethylene group	3.507	2.362	-1.145	5.9
24	7	6	More than one chlorine	4.158	1.012	-3.146	7.3
27	0	2	Iodine	-4.414	3.316	7.730	9.5

TABLE 15. TERATOGEN ALGORITHM EQUATION (Cont.)

KEYS	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	NON-TERATOGENS	TERATOGENS		NON-TERATOGENS	TERATOGENS	DIFFERENCE TER.-NONTER.	F Value
28	20	10	One -NH- group	2.355	0.689	-1.666	5.5
29	5	8	More than one -NH- group	1.093	3.079	1.986	3.0
31	3	4	More than one $-NH_2$ group	3.821	1.404	-2.417	3.7
38	13	4	More than one -OH group	3.848	1.139	-2.455	9.3
39	17	27	One -C=O group	0.679	1.902	1.223	3.9
41	18	13	One -C-OH (acid) group	4.978	2.553	-2.425	14.3
43	6	10	One -C-O (ester) group	-0.047	2.407	2.454	8.6
SUBSTITUENT FRAGMENTS							
K50	19	40	Generic halogen	0.182	1.872	1.690	8.6
52	11	9	More than one chlorine	3.492	1.517	-1.975	3.4
55	0	2	Iodine	-5.487	1.371	6.858	7.7
72	1	7	More than one -C (ester) group	-2.590	0.628	3.218	6.7

TABLE 15. TERATOGEN ALGORITHM EQUATION (Cont.)

KEY	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	NON-TERATOGENS	TERATOGENS		NON-TERATOGENS	TERATOGENS	DIFFERENCE TER.-NONTER.	F Value
RING HETEROATOMS							
K73	23	28	Single occurrence of oxygen	-0.873	0.198	1.071	2.5
77	34	25	Single occurrence of nitrogen	-0.335	0.949	1.284	4.3
78	31	47	Multiple occurrence of nitrogen	0.198	1.991	1.793	6.5
86	13	19	Multiple occurrence of carbonyl	1.892	0.505	-1.387	2.3
88	1	3	Multiple occurrence of carbonyl in more than one ring	-1.235	3.874	5.109	8.5
RING TYPES							
K99	29	57	Carbocyclic 6-membered ring	2.173	4.330	2.157	21.4
101	54	56	Heterocyclic 5-membered ring	1.594	0.079	-1.515	7.3
104	38	32	1 heteroatom in one ring	0.641	-0.285	-0.926	2.5
105	40	41	2 heteroatoms in one ring	1.555	-0.417	-1.972	9.9

TABLE 15. TERATOGEN ALGORITHM EQUATION (Cont.)

KEY	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	NON-TERATOGENS	TERATOGENS		NON-TERATOGENS	TERATOGENS	DIFFERINCE TER.-NONTER.	F VALUE
109	2	3	More than 2 heteroatoms in more than one ring	-2.313	0.053	2.366	2.4
RING FUSIONS							
K110	40	46	1 single heterocyclic ring	2.645	0.595	-2.050	8.7
111	16	19	More than 1 single Heterocyclic ring	4.318	1.761	-2.557	8.8
112	14	4	1 single carbocyclic ring	2.983	-1.482	-4.465	25.1
114	6	1	1 carbo/carbo fusion	0.908	-2.620	-3.528	6.3
118	25	20	1 carbo/hetero fusion	2.569	-1.178	-3.747	29.1
122	19	16	1 hetero/hetero fusion	3.282	1.900	-1.382	3.1
RING LINKAGES							
K127	8	7	True bridge indicator	3.000	-0.118	-3.118	8.4
128	3	2	1 multi-cyclic point	2.9933	-0.490	-3.423	4.5
UNUSUAL CONDITIONS							
K133	5	5	Inorganics	9.241	4.454	-4.787	12.7

TABLE 15. TERATOGEN ALGORITHM EQUATION (Cont.)

KEY	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	NON-TERATOGENS	TERATOGENS		NON-TERATOGENS	TERATOGENS	DIFFERENCE TER.-NONTER.	F VALUE
TOTAL RING FEATURES							
K135	29	35	2 ring system	-0.738	0.795	1.533	4.4
139	0	4	More than 2 benzene rings	-1.537	2.066	3.603	4.8
141	18	13	2 carbocyclic rings	1.982	-0.585	-2.567	10.8
143	53	56	1 heterocyclic ring	0.297	2.954	2.657	11.0
144	33	37	2 heterocyclic rings	0.680	1.860	1.1180	3.1
ADDITIONAL CHAIN FRAGMENTS							
K151	3	3	Chain secondary amide	-1.637	1.117	2.754	3.1
152	1	5	Chain tertiary amide	-0.277	3.013	3.290	5.7
159	0	4	Chain azo and diazo	-2.297	1.119	3.416	3.4
160	1	1	Chain C-nitroso	4.745	-0.868	-5.613	5.4
167	2	1	Chain methoxy	2.950	-3.573	-6.523	10.2
172	5	3	Chain phenoxy	2.769	0.958	-1.811	2.0
173	0	1	Chain phenylazo, and phenylhydrazo	0.811	5.464	4.653	2.3

TABLE 15. TERATOGEN ALGORITHM EQUATION (Cont.)

KEY	NUMBER OF OCCURRENCES		DESCRIPTION	COEFFICIENTS FOR GROUP			
	NON-TERATOGENS	TERATOGENS		NON-TERATOGENS	TERATOGENS	DIFFERENCE TER.-NONTER.	F Value
ADDITIONAL SUBSTITUENT FRAGMENTS							
K184	1	0	Substituent N-unsubstituted acylhydrazide	5.233	-1.107	-6.340	3.6
185	3	0	Substituent N-substituted acylhydrazides	3.761	-6.108	-9.869	18.4
188	5	13	Barbiturate	-2.524	1.477	4.001	10.2
193	9	5	Substituent sulfonamide	3.762	-0.755	-4.517	16.3
198	9	15	Substituent methoxy	-3.083	1.610	4.693	29.3
208	1	0	Substituent sulfamido	4.923	0.152	-4.771	2.2
307	0	1	Substituent other dialkylamino	-0.521	-7.378	-6.857	3.3
ADDITIONAL METAL FRAGMENTS							
K222	0	2	Cd	-2.053	3.353	5.406	5.6
251	0	2	Li	-4.414	3.316	7.730	9.5
256	0	4	Hg	4.554	7.000	2.446	2.4

From Enslein et al, (3).

TABLE 16. DISTRIBUTION OF TERATOGENICITY SCORES

N = 654

Score	N	%
0	9	1.4
.05	55	8.4
.10	58	8.9
.15	24	3.7
.20	32	4.9
.25	20	3.1
.30	14	2.1
.35	17	2.6
.40	28	4.3
.45	21	3.2
.50	23	3.5
.55	15	2.3
.60	18	2.8
.65	34	5.2
.70	49	7.5
.75	45	6.9
.80	60	9.2
.85	85	13.0
.90	40	6.1
.95	5	0.8
1.00	2	0.3

From Enslein et al, (3)

TABLE 17. TERATOGEN ALGORITHM - DISCRIMINANT EQUATION EVALUATION

		Teratogen		Indeterminate		Nonteratogen	
Indeterminate: 0.400 - 0.599		N	%	N	%	N	%
"ACTUAL" CLASSIFICATION	Teratogen	186	80.2	17	7.3	29	12.5
	Nonteratogen	27	13.9	19	9.8	148	76.3
Indeterminate: 0.300 - 0.699							
"ACTUAL" CLASSIFICATION	Teratogen	161	69.4	49	21.1	22	9.5
	Nonteratogen	15	7.7	46	23.7	133	68.6

From Enslein et al, (3)

TABLE 18. TERATOGEN ALGORITHM - MISCLASSIFICATION IN RANGES

Probability of Teratogenicity	No. of Compounds in range	Proportion Misclassified	Misclassifications N	Misclassifications Actual Cumulative	Misclassifications Expected Cumulative
.9 - 1.0	113	.009	1	1	5.65
.8 - .899	28	.179	5	6	9.85
.7 - .799	36	.278	10	16	18.85
.6 - .699	37	.324	12	28	31.8
.5 - .599	12	.583	7	35	37.2
.4 - .499	24	.500	12	47	48.0
.3 - .399	22	.318	7	54	55.7
.2 - .299	28	.464	13	67	62.7
.1 - .199	47	.170	8	75	69.75
.0 - .099	83	.048	4	79	73.9

From Enslein et al, (3)

--- Key generation program could not handle compound.

--- No keys generated for compound appeared in the respective estimation equation(s).

--- Compound was inorganic and, therefore, a sufficient number of keys for its molecular description could not be generated.

--- Molecular weight was not available (applies only to LD_{50} and Carcinogenesis Model, and only for very few compounds).

Of the approximately 8,000 chemicals listed in the NOHS data base, the following number of compounds met the algorithm requirements and an estimate was calculated for the toxic endpoint (3) indicated:

Model	No. of Compounds
LD_{50} (Oral Rat)	1937
Mutagenesis	2601
Carcinogenesis	2685
Teratogenesis	2338

A list of these compounds in ascending CAS number order are presented by toxic endpoint in Appendices M-P. Note that these lists of compounds are not the same as those listed in the modeling data bases. However, the majority of the modeling data base compounds are included in their respective NOHS data base listings by toxic endpoint.

The distribution of the number of compounds with toxicity estimate values of 0.750 or higher for each of the four toxic endpoints are shown in Table 19. Some general observations may be made as to the number and distribution of compounds in the upper ranges (0.750) of toxicity estimate values generated by the mutagen, carcinogen, and teratogen algorithms. The cut-off point values for the ranges selected are empirical and are used only as a convenient selection mechanism. Further refinement of the algorithms may permit a more subjective analysis. The user should not attach undue significance to groupings based only on empirical cut points. This is because a compound with a predicted endpoint value of 0.749 may actually have a true value of 0.800 when the standard deviation of the estimate is considered.

Although the carcinogen, mutagen, and teratogen algorithms were applied to over 2,300 compounds (mutagen - 2,601, carcinogen - 2,685, teratogen - 2,338) there is a considerable difference in the number of compounds scored as 1.000 between algorithms. The number of compounds for each algorithm with a predicted value of 1.000 (a perfect score) are 57, 153, and 25 for the mutagen, carcinogen, and teratogen algorithms respectively. The smaller number of compounds in the teratogen algorithm (25 compared to 153 for the carcinogen and 57 for the mutagen) may be partly due to teratology being a relatively new field of toxicology. Therefore teratogenicity has been the subject of research, testing, and information processing for a much shorter period of time than either

TABLE 19. NUMBER OF CHEMICAL COMPOUNDS BY SELECTED RANGES (0.750 OR GREATER) OF ESTIMATED TOXICITY ENDPOINT VALUES

Algorithm:	Mutagen	Carcinogen	Teratogen
* Total Number of compounds submitted	2601	2685	2338
Range:			
1.0	57	153	25
.990 - 1.0	185	424	131
.975 - 1.0	253	563	285
.950 - 1.0	312	640	447
.900 - 1.0	373	712	575
.750 - 1.0	551	905	847

carcinogenicity or mutagenicity. This may not be an appropriate explanation for the large difference between the number of perfect score compounds in the carcinogen and mutagen algorithms. However, it must be remembered that the mutagenicity modeling data base was constructed from compounds only on the basis of reported Ames test data for mutagenicity. There are a number of other tests for mutagenicity that were not considered in developing this algorithm (26), which may have had significant effects on the mutagen model output.

One way of expressing the degree of acute toxicity of the more toxic compounds is to realize that ratios of 1:100,000 and 1:1,000 are represented by the doses of 10 mg/kg and 1g/kg respectively. Of the 1,937 compounds receiving LD_{50} estimate values, 47 fell in the 0.0 to 10 mg/kg dose range and 1,046 fell in the 0.0 to 1g/kg dose range.

The incorrect values estimated for some of the compounds in each algorithm are indicative of the degree of statistical error unavoidably associated with these algorithms. For example, sucrose is estimated to have a carcinogenic potential of 0.944, and a number of other sugars also have high carcinogenic values. These high estimate values are due to the occurrence of a six-membered heterocyclic ring in the structure of these compounds. This substructure (represented by key 102) has a high coefficient value (definite score) of 7.20 due to the presence of this generic substructure in some highly carcinogenic six-membered heterocyclic ring compounds. Subsequent development of this algorithm will address other identified inadequacies of this algorithm also known to cause false high estimate values (3).

Since the contractual completion of the four algorithms, they have been tested against compounds not used in the modeling process for which toxicity data had recently become available. The results of this testing provide a means to assess the ability of the algorithms to discriminate toxic endpoints (yes or no) or to predict LD_{50} values. The LD_{50} algorithm was applied to 908 compounds for which reported LD_{50} (oral, rat) data had recently become available. The algorithm predicted an LD_{50} estimate that was within a range of 0.4X to 2.5X for 50% of the compounds and 0.2x to 5x for 80% of the compounds (27).

The mutagen algorithm was applied to 60 compounds for which reported mutagenicity data had become available since development of the algorithm. Of these 60 compounds, 50 received an estimate of mutagenicity with the remaining 10 receiving indeterminant (i.e., no decision) estimates. The algorithm correctly classified 41 of 50 (82%) compounds as mutagenic or nonmutagenic and misclassified 9 of 50 (18%) compounds (27).

The carcinogen algorithm was applied to 78 compounds for which reported carcinogenicity data had become available since the algorithm was developed. Of these, 38 compounds received estimates of carcinogenicity, however 3 of these were indeterminate estimates. The algorithm correctly classified 25 of 35 (71.4%) compounds as carcinogenic and misclassified 10 of 35 (28.6%). Enslein et al, (27) noted that steroid compounds generally exert their influence through indirect means via the endocrine system, therefore structure-activity relationship concepts would not be as applicable as they are for

compounds with more direct acting toxic effects. Removing the steroid compounds left 23 compounds receiving an estimate of carcinogenicity. The algorithm correctly classified 19 of 23 compounds (82.5%) and incorrectly classified 4 of 23 compounds (17.5%).

The teratogen algorithm was applied to only 5 compounds for which reported teratogenicity data had become available. The algorithm correctly classified 4 of 5 compounds as teratogenic. Enslein (27) notes that 3 of the 5 compounds were tested in only one species. These results show that the algorithms are able to discriminate and predict toxic endpoints with a reasonable degree of reliability.

IV. Discussion

The concept of predicting chemical activity from chemical structure, more commonly referred to as structure-activity relationships (SARs), has been and is used in several areas of chemistry and related fields (e.g., biochemistry, pharamacology, and toxicology). The development and computerized application of SARs has allowed considerable advances in both the accuracy and the specificity of SAR models. As exemplified in this project, SAR concepts linked with computer processing represent a valuable tool for toxicologic research. A number of other SAR studies regarding predictive toxicology have been reported, some of which are listed in Table 20 and will be briefly discussed here.

Van Duuren has published results indicating that "...studies show that this approach of structure-activity studies can lead to prediction of human carcinogenicity before epidemiologic studies are carried out." He has worked primarily with halogenated hydrocarbons. (28, 29)

Computer assisted structure-activity studies on nitrosamine compounds have been reported by Chou and Jurs (30). The pattern-recognition approach they developed is reported to have a predictive ability of 91%-93% overall for carcinogens and 85% for noncarcinogens in separating 116 carcinogens from 28 noncarcinogens. From this they concluded that "This relatively high predictability demonstrates that pattern-recognition methods can be useful in analyzing these compounds for carcinogenic activity "(30).

In similar studies, Yuan and Jurs reported on computer-assisted structure-activities of polycyclic aromatic hydrocarbons (PAHs). They were able to classify 191 PAHs as carcinogens or noncarcinogens with a predictive ability of 90.5% correct classification (31).

In 1978, Spann et al, reported on "the possibility of a computer program to predict probable metabolites of a compound using metabolic reactants and a specific set of reactions" (32). The metabolism of some compounds must be considered especially in toxicology, because the structure of the metabolite rather than that of the parent compound may be responsible for the observed biological effect (32). Applying such a model to selected toxic compounds, e.g., carcinogens, could produce new insight into the mechanism(s) of toxicity. Taking the usefulness of this approach one step further, the potential for investigating chemical synergism, especially in the field of toxicology, may produce significant advances in understanding the mechanisms involved.

TABLE 20. SOME PREDICTIVE TOXICOLOGY ORIENTED MODELS FOR THE CORRELATION OF CHEMICAL STRUCTURE WITH A BIOLOGIC ENDPOINT

Chemical(s)	Model Type	Researchers	Ref. No.
Halogenated hydrocarbons	SAR - Carcinogen	Van Duuren	28, 29
Catecholamines	SAR - Metabolities	Spann et al	32
Nitrosamines	SAR	Chou et al	30
PAH	SAR	Yuan et al	31
Organochlorines	Partition Coefficient	Freed et al	35
Organophosphates	Partition Coefficient	Freed et al	35
Teratogens - diverse	Review for structural diverse correlation	Dyban	34
Mutagens - diverse	QSAR	Johnson et al	11
PAH	Quantum chemical	Lehr et al	11
Diverse - LD_{50} (Rat Oral)	SAR - Model	Enslein et al	3
Diverse - Carcinogen	SAR - Model	Enslein et al	3
Diverse - Mutagen (Ames Test)	SAR - Model	Enslein et al	3
Diverse - Teratogen	SAR - Model	Enslein et al	3

The investigation of teratogenic compounds by the structure-activity approach has been advocated by Dyban (33). He states that in view of the already large and rapidly growing number of chemicals in the environment that it would not be practical to test for teratogenicity using current routine methods. Instead, it would be more productive to improve the prediction of teratogenic chemicals in order to study the mechanisms involved. As a demonstration of the close association between teratogenic activity and specific features of chemical structure, Dyban cites evidence of teratogenic activity regularly fluctuating as a function of molecular structure (33).

The need for quick and reliable tests to assess the toxicity of existing or new compounds is widely agreed upon. How to derive and implement such tests is the subject of considerable disagreement. In his paper "Criteria for Selecting Chemical Compounds for Carcinogenicity Testing: An Essay" (34), Arcos lists four categories of selection criteria:

1. structural criterion
2. operational criterion (complementary to structural criterion)
3. "guilt" by association criterion
4. "after the fact" criterion

Although he contends that molecular structure alone will not provide the sole basis for adequately assessing carcinogenic potential (reference criterion No. 2) it is apparent that it does play a major role in such an assessment.

Aside from the well known limitations associated with extrapolating animal toxicity data to humans, the four algorithms developed in this project have additional limitations. Toxicity estimates generated by the algorithms for mutagenicity, carcinogenicity, and teratogenicity are relative estimates. For example, the teratogen algorithm provides an estimate of teratogenicity on a scale of 0.000 to 1.000, relative to a known teratogenic compound such as thalidomide which has a value of 1.000. Such estimates are not absolute in nature because they are derived by an imperfect model developed from a data base that, as a subset of the "chemical universe", is undoubtedly a biased representation (3).

The validity and reliability of statistical values, such as regression constants and coefficient values assigned to the chemical descriptor keys generated for these four algorithms pertain only to those compounds in the respective modeling data base. Such values would be expected to change, perhaps markedly, when these algorithms are applied to compounds not in the modeling data bases. This is an unavoidable complication in using predictive algorithms at such an embryonic level in the stages of algorithm development and application. As the number and quality of testing of compounds available for inclusion in the modeling data bases increase more efficient algorithms should be possible, and other approaches for modeling may prove to be more effective in predicting toxic endpoints.

Although the algorithms do not directly address the problem of metabolism of parent compounds into other more or less toxic compounds, they do so indirectly because the toxic effects that are observed and

reported are used as the basis for assigning a toxic or non-toxic status of the parent compound. Should individual metabolites be identified, they too could receive estimates of toxicity. The purpose of the algorithms is not to investigate toxicokinetics but rather to discriminate toxic potential or estimate LD_{50} concentrations. It should also be pointed out that the algorithms do not consider site of toxic effect, e.g., type of cancer, or the actual cause of death in the case of the LD_{50} algorithm. Enslein et al, have proposed that these and other considerations be included in future modeling efforts (3).

These models serve only as tools for investigating the potential toxicity of compounds. Therefore, it must be emphasized that, because these algorithms produce only estimate values of toxicity, such values should not be used as a sole basis for declaring a compound as being toxic or nontoxic for the endpoint in question. Further, these values should not be used in place of animal test data. Instead these algorithms should only be used to provide a quick estimate of toxic potential, and may be incorporated into a larger process of ranking or prioritizing compounds as regards toxic potential.

REFERENCES

1. National Occupational Hazard Survey, National Institute for Occupational Safety and Health, DHEW Publication No. (NIOSH) 74-127, May, 1974; (NIOSH) 77-213, July, 1977; and (NIOSH) 78-114, December, 1977.

2. Chemical and Engineering News, March 9, p. 26, 1981.

3. Enslein, K.: Four statistical structure-activity models for the prediction of toxicological endpoints. Draft report on NIOSH contract no. 210-80-0044, June, 1981.

4. Adamson, G.W., and Bawden, D.: A method of structure - activity correlation using Wiswesser Line-Formula Notation. J. Chem. Info. Comput. Sci., Vol. 15, 4: 215-220, 1975.

5. Jurs, P.C. and Chou, J.T.: Computer-assisted computation of partition coefficients from molecular structures using fragments. J. Med. Chem. 22: 476-483, 1979.

6. Chou, J.T. and Jurs, P.C.: Computer-assisted studies of chemical carcinogens - a heterogenous data set. J. Chem. Info. Comput. Sci. 19: 172-178, 1979.

7. Smith, I.A. et al: Relationships between carcinogenicity and theoretical reactivity indices in polycyclic aromatic hydrocarbons. Cancer Res. 38: 2968-2977.

8. Purcell, W.P. et al: Strategy of drug design - a guide to biological activity. Wiley, New York, 1973.

9. Kaufman, J.J.: Quantum chemical and physiochemical influences on structure-activity and drug design. Int. J. Quantum Chem. 16: 221-241, 1979.

10. Loew, G. et al: Quantum chemical studies of polycyclic aromatic hydrocarbons and their metabolites: correlations to their carcinogenicity. Chem. - Biol. Interact 26: 75 - 89, 1979.

11. Asher, I.M. and Zervos, C. (Eds): Symposium on structural correlates of carcinogenesis and mutagenesis: a guide to testing priorities? Proceeding of the second FDA Office of Science Summer Symposium, 1977.

12. Craig, P.N. and Waite, J.H.: Analysis and trial application of correlation methodologies for predicting toxicity of organic chemicals. Natl. Tech. Inf. Serv. (NTIS) PB-258, 119/7GA, 1976.

13. Enslein, K. and Craig, P.N.: A toxicity estimation model. J. Envir. Path Toxicol. 2: 115-121, 1978.

14. Smith, E.G. and Baker, P.A.: The Wiswesser Line-Formula Chemical Notation (WLN). (third edition) Chemical Information Management, Inc. (CIMI), Cherry Hill, NJ, 1975.

15. Koniver, D.A., Wiswesser, W.J., and Usdin, E.: Wiswesser Line-Formula Notation: Simplified techniques for converting chemical structures to WLN. Science, Vol. 170, 30 June 1972.

16. Enslein, K., Wilf, H. S., Ralston, A. (Eds.) Statistical Methods for Digital Computers, Chapter Four. Wiley, New York, New York, 1977.

17. Hocking, R. R., The analysis and selection of variables in linear regression. Biometrics 32: 1-49, 1976.

18. Siegel, S.: Nonparametric Statistics for the Behavioral Sciences. McGraw - Hill, New York, 127-136, 1956.

19. Marguardt, D. W. and Snee, R. D. Ridge regression in practice. Amer. Statistician 29: 3-20, 1975.

20. Christensen, H. E., editor. The Toxic Substances List (Now RTECS). HEW Publication No. (NIOSH) 74-134, 1975.

21. Waters, W. D. and Auletta, A.: The GENE-TOX Program: Genetic activity evaluation. J. Chem. Inf. Comput. Sci. 21: 35-38, 1981.

22. Craig, P.N.: Personal Communication, February, 1983.

23. IARC Working Group (1980): Chemicals and industrial processes associated with cancer in humans; and evaluation of human evidence and animal data, IARC monographs, Vol. 1-20. Cancer Res. 40: 1-12,1980.

24. Shepard, I.H.: Catalog of Teratogenic Agents, 3rd Ed. Johns Hopkins Univ. Press, Baltimore, MD, 1980.

25. Schardein, J.L.: Drugs as Teratogens. CRC Press, Cleveland, OH, 1976.

26. Handbook of Mutagenicity Test Procedures. Kirbey, Legator, Nichols, and Ramel (editors), Elsevier Scientific Publishing Company, New York - Oxford, 1977.

27. Enslein K. and Craig, P.N.: Validation of a set of predictive structure-activity models of toxicological endpoints. Revised abstract of a paper presented at the National ACS Meeting, Kansas City, Missouri, September 14, 1982.

28. Van Duuren, B.L.: Chemical Structure, reactivity, and carcinogenicity of halohydrocarbons. Env. Hlth Persp. 21: 17-23, 1977.

29. Van Duuren, B.L., Katz, C., Goldschmidt, B.M., Frenkel, K., and Sivak, A.: Carcinogenicity of halo-ethers. II.: Structure-activity relationships of bis(chloromethyl) ether. J. Natl. Cancer Inst. 48 (5): 1432-1439, 1972.

30. Chou, T.J. and Jurs, P.C.: Computer-assisted structure-activity studies of chemical carcinogens: An N-nitroso compound data set. J. Med. Chem. 22(7): 792-797,1979.

31. Yuan, M. and Jurs, P.C.: Computer-assisted structure-activity studies of chemical carcinogens: A polycyclic aromatic hydrocarbon data set. Tox. App. Pharm. 52: 294-312, 1980.

32. Spann, M.L., Chu, K.C., Wipke, W.T., and Ouchi, C.: Use of computerized methods to predict metabolic pathways and metabolites. J. Env. Path. Tox. 2: 123-131, 1978.

33. Dyban, A.P.: Topical problems and basic developmental trends of investigations concerning the embryotoxic and teratogenic effect of environmental chemicals. Env. Hlth. Per. 30: 99-103, 1979.

34. Arcos J.C.: Criteria for selecting chemical compounds for carcinogenicity testing: and essay. J. Env. Path. Tox. 1: 433-458.1978.

35. Freed, V. H., Chou, C.T., Schmeddling, D., and Kohnert, R.: Some physical factors in toxicological assessment tests. Env. Hlth. Per. 30: 75-80, 1979.

APPENDIX A
WISWESSER LINE-FORMULA NOTATION SYMBOLS AND DEFINITIONS

All of the international atomic symbols are used except K, U, V, W, Y, Cl, and Br. Two-letter stomic symbols in organic notations are enclosed between hyphens. Single letters preceded by a blank space indicate ring positions. Single letters not preceded by a blank space have the following meanings:

A generic alkyl
B boron atom
C unbranched carbon atom multiply bonded to an atom other than carbon or doubly bonded to two other carbon atoms
D symbol for chelate bond and initial symbol of a chelate notation
E bromine atom
F fluorine atom
G chlorine atom
H when preceded by a locant within ring signs, shows the position of a carbon atom bonded to four other atoms; elsewhere H means hydrogen atom.
I iodine atom
J sign for the end of a ring description
K nitrogen atom bonded to more than three atoms
L first symbol of a carbocyclic ring notation
M imino or imido -NH- group
N nitrogen atom, hydrogen free, attached to no more than three other atoms
O oxygen atom, hydrogen free
P phosphorus atom
Q hydroxyl group, -OH
R benzene ring
S sulfur atom
T first symbol of a heterocyclic ring notational; or within ring signs indicates a ring containing two or more carbon atoms each bonded to four other carbons
U double bond
V carbonyl connective, C=O (carbon attached to three other atoms)
W nonlinear (branching) dioxo group (as in $-NO_2$ or $-SO_2-$)
X carbon atom attached to four atoms other than hydrogen
Y carbon atom attached to three atoms other than hydrogen or doubly bonded oxygen
Z $-NH_2$ group
& punctuation mark showing the end of a side chain; or preceded by a space, sign of ionic salt, addition compound or suffixed information; or within ring signs indicates a ring NOT containing two or more carbon atoms that are bonded to four other atoms; or following a hyphen, shows certain spiro ring connections
\- separator or connective or other special uses
/ precedes each nonconsecutive locant pair; encloses polymer notations
* (1) points of attachments in polymer repeat units
(2) coincident atoms in polymer notations
(3) a multiplier symbol in inorganic notations
. space-filling symbol for inorganic notations

Numerals preceded by a space - are multipliers of preceding notation suffixes; or within rings signs L...J, T...J, or D...J show the number of multicyclic points in the ring structure.

Numerals not preceded by a space show ring sizes if within the ring signs; elsewhere numerals show the length of internally saturated, unbranched alkyl chains and segments.

Letters following a space and hyphen are proposed as symbols with special meanings to denote stereoisomerism.

From: Smith and Baker (14).

APPENDIX B
WLN EXAMPLE FOR AN ACYCLIC COMPOUND

ACYCLIC COMPOUND:

1. Compound: 2-chloro-1, 3-butadiene (chloroprene)

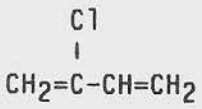
2. RTECS No.: EI9625000

3. CAS No.: 126998

4. Molecular Formula: C_4H_5Cl

5. Molecular Weight: 88.54

6. Molecular Structure:



7. WLN: 1UYG1U1

8. Explanation of WLN symbols:

   1= in this case, shows the length of internally saturated, unbranched alkyl chains and segments.

   U= double bond

   Y= carbon atom attached to three atoms other than hydrogen or a doubly bonded oxygen.

   G= chlorine atom

   1= see previous explanation

   U= see previous explanation

NOTE: In general, symbols are assigned in a linear fashion starting at one end of the molecule and proceeding to the other end.

APPENDIX C
WLN EXAMPLE FOR A CYCLIC COMPOUND

Cyclic compound:

1. Compound: 4,4'-stilbenediol, alpha, alpha', -diethyl (diethyl stilbesterol).

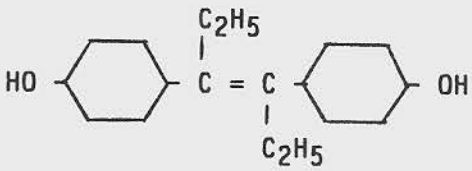
2. RTECS No: WJ5600000.

3. CAS No.: 2698411.

4. Molecular Formula: $C_{18}H_{20}O_2$.

5. Molecular Weight: 268.38

6. Molecular Structure:



7. WLN: QR DY2&UY2&R DQ

8. Explanation of WLN symbols:

   Q = Hydroxyl group

   R = Benzene ring

   Space preceding a letter indicates location on ring; in this case the hydroxyl group is located at the D position of the benzene ring

   D = see preceding definition

   Y = Carbon atom attached to three atoms other than hydrogen or doubly bonded oxygen

   2 = in this case the numeral shows the length of one internally saturated, unbranched alkyl chains and segments

   & = in this case it shows the end of a side chain

   U = double bond

   Y = see previous explanation

   2 = see previous explanation

   & = see previous explanation

R = see previous explanation

Cyclic compound: (continued)

Space - see previous explanation

D = see previous explanation

Q = see previous explanation

NOTE: In general, symbols are assigned in a linear fashion starting at one end of the molecule and proceeding to the other end.

APPENDIX D
MOLECULAR SUBSTRUCTURE KEYS AND THEIR DEFINITIONS

Keys 1 to 336 were used in development of the carcinogenesis model. Keys 1 to 350 (excluding 335 and 336) were used in development of the teratogenesis, mutagenesis and LD_{50} models. The following special symbols have been used here in the WLN descriptions:

- (underscore)	- a character string may intervene
&	- a terminal substituent
^	- single space (Note: symbol will not appear in the actual WLN)
(C)	- any carbon atom (e.g. X,Y,1,2,etc.)
(N)	- any nitrogen atom (e.g. K,M,N,Z)
9	- any numeric

ALL PARTS OF THE MOLECULE

(1) Atoms other than C,H,O,N,S or halogens - Character sequence -aa- or the character B (not ^ B) or the character P (not ^ P) found anywhere in the molecule, or the sequence -E-, -F-, -G-, -I- found in a ring.

(2) Positive charge - Character sequence ^ &Q^ indicating quarternary salt present, at end of true WLN notation.

ALL NON-CYCLIC PARTS OF THE MOLECULE

Character sequences must be outside ring signs.

(3) Branching terminal nitro group (NO2) - The character sequence NW (or WN at the start of the notation).

(4) Dioxo (excluding NO2) The character sequence W but not NW or WN. Any substituent W found within ring signs is also included here.

(5) Terminal oxygen (not carbonyl) - The character sequence O& or O^, or the letter O starting the notation.

(6) One 3-branch carbon atom - The character Y (but not ^ Y) occurring once only. (Note: More than one 3-branch carbon is fragment 148.

(7) 4-branch carbon atom - The character X (but not ^ X)

(8) 3-branch nitrogen atom - The character N, but not N or NW or WN or NU or UN. This definition also includes unusual conditions of nitrogen, e.g. in cyanide, isocyanide, etc.

(9) Greater than 3-branch nitrogen atom - The character K but not ^K.

(10) 1 sulphur atom - The single occurrence of S, but not ^S or US^ or US&.

(11) More than 1 sulphur atom - The multiple occurrence of S, but not ^S or US^ or SU or US&.

(12) 1 -C=S group - The single occurrence of the groups US^ or ^US& (or SU at the start of the notation only).

(13) More that 1 -C=S group - The multiple occurrence of the groups US^ or US& (or SU at the start of the notation only).

(14) 1 double bond, excluding -C=S, or -C=U - The single occurrence of the letter U, but not in any of the following groups, ^U, UU, US^ , US&, SU, NU, UN, MU, UM.

(15) More than 1 double bond, excluding -C=S, -N=, or -C=O - The multiple occurrence of the letter U, but not in any of the following groups, ^U, UU, US^ , US&, SU, UN, NU, MU, or UM.

(16) Triple Bond - The occurrence of the symbol combination UU.

CHAIN FRAGMENTS

Character sequences must not be immediately attached to a ring system.

(17) 1 methyl/methylene group - Single occurrence of the number 1 not followed or preceded by a numeral.

(18) More than 1 methyl/methylene group - Multiple occurrence of the number 1 not followed or preceded by a numeral.

(19) Ethyl/ethylene group - Occurrence of the number 2 not followed by or preceded by a numeral.

(20) Alkyl chain (CH2)n or CH3(CH2)n-1 where n=3 to 9 - Occurrence of a number in the range 3-9, but not followed by or preceded by a numeral.

(21) Alkyl chain (CH2)n or CH3(CH2)n-1 where n=10 or more - Occurrence of a number in the range of 10 or more, but not followed by or preceded by a numeral.

(22) Generic halogen - Occurrence of any of the characters E,F,G, or I.

(23) One chlorine - Single occurrence of the character G.

(24) More than one chlorine - Multiple occurrence of the character G.

(25) Bromine - Occurrence of one or more E symbols.

(26) Fluorine - Occurrence of one or more F symbols.

(27) Iodine - Occurrence of one or more I symbols.

(28) One -NH- group - Single occurrence of the symbol M, but not UM (or MU at the start of the notation).

(29) More than one -NH- group - Multiple occurrence of the symbol M, but not UM (or MU at the start of the notation).

(30) One -NH2 group - Single occurrence of the symbol Z.

(31) More than one -NH2 group - Multiple occurrence of the symbol Z.

(32) One -N= or HN= group - Single occurrence of the symbol sequence UN or NU or UM (or MU) at the start of the notation.

(33) More than one -N= or HN= group - Multiple occurrence of the symbol sequence UN or NU or UM (or MU at the start of the notation.

(34) Unusual carbon atom - One or more occurrences of the symbol C. Usually found in triple bonds, such as cyanides, isocyanides, etc.

(35) One -O- group - Single occurrence of the symbol O, but not in the sequence OV or VO, or as O^ or O&.

(36) More than one -O- group - More than one occurrence of the symbol O, but not in the sequence VO or OV or O& or O^.

(37) One -OH group - Single occurrence of the symbol Q, but not in the sequence VQ (or QV at the start of the notation).

(38) More than one -OH group - Multiple occurrence of the symbol Q, but not in the sequence VQ (or QV at the start of the notation).

(39) One -C=O group - Single occurrence of the symbol V, but not in the sequence VQ or VO or OV (or QV at the start of the notation).

(40) More than one -C=O group - Multiple occurrence of the symbol V, but not in the sequence VQ or VO or OV (or QV at the start of the notation).

(41) One -C(=O)-OH (acid) group - Single occurrence of the symbol combination VQ (or QV at the start of the notation).

(42) More than one -C(=O)-OH (acid) group - Multiple occurrence of the symbol combination VQ (or QV at the start of the notation).

(43) One -C(=O) -O (ester) group - Single occurrence of the symbol combination VO or OV.

(44) More than one -C(=O) -O (ester) group - Multiple occurrence of the symbol combination VO or OV.

ADDITIONAL CHAIN FRAGEMENTS

(150) Chain primary amide - Character sequence ZV or VZ bonded to acyclic C only.

(151) Chain secondary amide - Character Sequence VM or MV bonded to acyclic C only

(152) Chain tertiary amide - Character sequence N_V or VN bonded to acyclic C only.

(153) Chain N-unsubstituted acylgydrazide - Character sequence ZMV or VMZ bonded to acyclic C.

(154) Chain N-Substituted acylhydrazides - Character sequence MMV, VMM, MN_V, VN_M, N_N_V, N_V(&)N, N_MV, VMN, ZN_V, or VNZ, with V bonded to acyclic C only.

(155) Chain primary amidine - Character sequence MUYZ or YZUM bonded to acylic C only.

(156) Chain amidine - Character sequence (N)_Y_UN, or NUY_(N) bonded to acyclic C only and excluding (key 155).

(159) Chain azo and diazo - Character sequence NUN, UNN, or NNU.

(160) Chain C-nitroso - Character sequence ON or NO bonded to acyclic C.

(161) Chain N-nitroso - Character sequence ON(N) or (N)_NO.

(162) Chain sulfonamide - Character sequence (N)_SW or SW(N), (excluding key 177).

(163) Chain guanidine - Character sequence (N)_Y(N)_U(N) or (N)UY_(N)_(N).

(164) Chain N-N, azoxy - Character sequence (N)-(N), not part of another key (e.g. 153,154), and NUNO& and NO&UN (excluding key 305).

(165) Chain thioamide - Character sequence SUY(N) or Y(N)_US.

(166) Chain dialkylamino - Character sequence 9N9 & (at beginning of notation) or N9&9 or N9&9&, bonded to carbon but excluding carbonyl and cyclic carbon.

(Note: Key 304 is chain dialkylamino, not bonded to carbon.)

(167) Chain methoxy - Character sequence O1 or 1O (O=letter), bonded to acyclic C, except carbonyl; methyl group must be terminal.

(168) Chain hydroxylamine - Character sequence Q(N) or (N)_Q.

(169) Chain oxime - Character sequence QNU or UNQ.

(170) Chain N-nitro - Character sequence WN(N) or (N)_NW.

(171) Chain phenethyl - Character sequence R2 or 2R.

(172) Chain phenoxy - Character sequence RO or OR.

(173) Chain phenylazo, and phenylhydrazo - Character sequence RMNU, RNUN, NUNR, or UNMR.

(174) Chain phenylureido - Character sequence RMUM and MUMR.

(175) Chain phosphonyl - Character sequence QPQO& or PQQO (where O is terminal), but excluding P attached to 4 O atoms.

(176) Chain semicarbazide and semicarbazone - Character sequence MMVZ, ZVMM, UNMVZ, or ZVMNU.

(177) Chain sulfamido - Character sequence MSWQ or WSQM.

(178) Chain urea - Character sequence (N)_V(N).

(179) Chain cyano - Character sequence NC or CN (where N is terminal).

(304) Chain other dialkylamino - Character sequence 9N9& (at beginning of notation) or N9&9 or N9&9&, not bonded to carbon or a ring.

(306) Chain carbamate - Character sequence OV(N), or (N)_VO.

SUBSTITUENT FRAGMENTS

The type of fragment in this class is exactly equivalent to the chain class of fragments. The fragments must be directly attached to a ring of some kind, and may be found after a locant in the notation or attached to a trailing ring system.

(45) One methyl/methylene group - Single occurrence of the number 1 not followed or preceded by a numeral.

(46) More than one methyl/methylene group - Multiple occurrence of the number 1 not followed or preceded by a numeral.

(47) Ethyl/ethylene group - Occurrence of the number 2 not followed by or preceded by a numeral.

(48) Alkyl chain (CH2)n or CH3(CH2)n-1 where n= 3-9 - Occurrence of a number in the range 3 to 9, but not followed by or preceded by a numeral.

(49) Alkyl chain (CH2)n or CH3(CH2)n-1 where n= 10 or more - Occurrence of a number in the range of 10 or more, but not followed by or preceded by a numeral.

(50) Generic halogen - Occurrence of any of the characters E,F,G,I.

(51) One chlorine - Single occurrence of the character G.

(52) More than one chlorine - Multiple occurrence of the character G.

(53) Bromine - Occurrence of one or more E symbols.

(54) Fluorine - Occurrence of one or more F symbols.

(55) Iodine - Occurrence of one or more I symbols.

(56) One -NH- group - Single occurrence of the symbol M, but not UM (or MU at the start of the notation).

(57) More than one -NH- group - Multiple occurrence of the symbol M, but not UM (or MU at the start of the notation.)

(58) One -NH2 group - Single occurrence of the symbol Z.

(59) More than one -NH2 group - Multiple occurrence of the symbol Z.

(60) One -N= or HN= group - Single occurrence of the symbol sequence UN or NU or UM (or MU at the start of the notation).

(61) More than one -N= or HN= group - Multiple occurrence of the symbol sequence UN or NU or UM (or MU at the start of the notation).

(62) Unusual carbon atom - One or more occurrences of the symbol C. Usually found in triple bonds, such as cyanides, isocyanides, etc.

(63) One -O- group - Single occurrence of the symbol O, but not in the sequence OV or VO or O ^ or O&.

(64) More than one -O- group - More than one occurrence of the symbol O, but not in the sequence VO or OV or O ^ or O&.

(65) One -OH group - Single occurrence of the symbol Q, but not in the sequence VQ (or QV at the start of the notation).

(66) More than one -OH group - Multiple occurrence of the symbol Q, but not in the sequence VQ (or QV at the start of the notation).

(67) One -C=O group - Single occurrence of the symbol V, but not in the sequence VQ or VO or OV (or QV at the start of the notation).

(68) More than one -C=O group - Multiple occurrence of the symbol V, but not in the sequence VQ or VO or OV (or QV at the start of the notation).

(69) One $-\overset{O}{\overset{\|}{C}}-OH$ (acid) group - Single occurrence of the symbol combination VQ (or QV at the start of the notation).

(70) More than one $-\overset{O}{\overset{\|}{C}}-OH$ (acid) group - Multiple occurrence of the symbol combination VQ (or QV at the start of the notation).

(71) One $-\overset{O}{\overset{\|}{C}}-O$ (ester) group - Single occurrence of the symbol combination VO or OV.

(72) More than one $-\overset{O}{\overset{\|}{C}}-O-$ (ester) group - Multiple occurrence of the symbol combination VO or OV.

ADDITONAL SUBSTITUENT FRAGMENTS

(180) Biphenyl - Character sequence R locR.

(181) Substituent primary amide - Character sequence ZV or VZ bonded to ring C only.

(182) Substituent secondary amide - Character sequence VM or MV bonded to ring C only.

(183) Substituent tertiary amide - Character sequence N_V or VN bonded to ring C only. Note that (N) includes N in a ring.

(184) Substituent N-unsubstituted acylhydrazide - Character sequence ZMV or VMZ bonded to ring C.

(185) Substituent N-substituted acylhydrazides - Character sequences MMV, VMM, MN_V, VN_M, N_V, N_V(&)N, N_MV, VMN, ZN_V, or VNZ, with V bonded to ring C only. Note that (N) includes N in a ring.

(186) Substituent primary amidine - Character sequence MUYZ or YZUM bonded to ring C only.

(187) Substituent amidine - Character sequence (N)_Y_UN, or NUY_(N) bonded to ring C only and excluding (key 155). Note that (N) includes N in a ring.

(188) Barbiturate - Character sequence (N)V(N)V or V(N)V(N) within ring symbols.

(189) Lactam - Character sequence (N)V or V(N) within ring symbols (excluding key 188).

(190) Substituent azo and diazo - Character sequence NUN, UNN, or NNU.

(191) Substituent C-nitroso - Character sequence ON or NO bonded to ring C.

(192) Substituent N-nitroso - Character sequence ON(N) or (N)_NO. Note that (N) includes N in a ring.

(193) Substituent sulfonamide - Character sequence (N)_SW or SW(N), (excluding key 177). Note that (N) includes N in a ring.

(194) Substituent guanidine - Character sequence (N)Y(N)_U(N) or (N)UY(N) U(N). Note that (N) includes N in a ring.

(195) Substituent N-N - Character sequence (N)_(N), not part of another key (e.g. 153, 154). Note that (N) includes N in a ring.

(196) Substituent thioamide - Character sequence SUY(N) or YN_US.

(197) Substituent dialkylamino - Character sequence 9N9& (at beginning of notation) or N9&9^ or N9&9&, bonded to cyclic carbon. (Note: Key 307 is substituent dialkylamino, not bonded to ring carbon).

(198) Substituent methoxy - Character sequence O1 or 1O (O=letter), bonded to acyclic C, except carbonyl; methyl group must be terminal.

(199) Substituent hydroxylamine - Character sequence Q(N) or (N)_Q. Note that (N) includes N in a ring.

(200) Substituent oxime - Character sequence QNU or UNQ.

(201) Substituent N-nitro - Character sequence WN(N) or (N)_NW. Note that (N) includes N in a ring.

(202) Substituent phenethyl - Character sequence R2 or 2R.

(203) Substituent phenoxy - Character Sequence RO or OR,

(204) Substituent phenylazo and phenylhydrazono - Character sequence RMNU, RNUN, NUNR, or UNMR.

(205) Substituent phenylureido - Character sequence RMUM or MUMR.

(206) Substituent phosphonyl - Character sequence QPQO& or PQQO (where O is terminal), But excluding P attached to 4 O atoms.

(207) Substituent semicarbazide and semicarbazone - Character sequence MMVZ, ZVMM, UNMVZ, or ZVMNU.

(208) Substituent sulfamido - Character sequence MSWQ or WSQM.

(209) Substituent ureas - Character sequence (N)_V(N). Note that (N) includes N in a ring.

(210) Substituent cyano - Character sequence NC or CN (where N is terminal).

(305) Aromatic azoxy - Character sequence NUNO& or NO&UN on benzene ring.

(307) Substituent other dialkylamino - Character sequence N9&9^ or N9&9&, bonded to a cyclic heteroatom.

(309) Substituent carbamate - Character sequence OV(N), or (N)_VO.

RING HETEROATOMS

Each ring system in the molecule is analyzed; each ring is isolated and assigned a heteroatomic description. This description lists the heteroatoms present in the ring. The following fragments are set according to the analysis of that ring description.

(73) Single occurrence of oxygen - A ring description contains only one oxygen (O).

(74) Multiple occurrence of oxygen - A ring description contains more than one oxygen.

(75) Single occurrence of oxygen in more than one ring - More than one ring description each containing only one oxygen.

(76) Multiple occurrence of oxygen in more than one ring - More than one ring description containing more than one oxygen.

(77) Single occurrence of nitrogen - A ring description contains one nitrogen (N, M, K).

(78) Multiple occurrence of nitrogen - A ring description contains more than one nitrogen.

(79) Single occurrence of nitrogen in more than one ring - More than one ring description containing one nitrogen.

(80) Multiple occurrence of nitrogen in more than one ring - More than one ring description containing more than one nitrogen.

(81) Single occurrence of sulphur - A ring description contains only one sulphur atom (S).

(82) Multiple occurrence of sulphur - A ring description contains more than one sulphur

(83) Single occurrence of sulphur in more than one ring - More than one ring description contains one sulphur.

(84) Multiple occurrence of sulphur in more than one ring - More than one ring description containing more than one sulphur.

(85) Single occurrence of carbonyl - A ring description contains one carbonyl (V).

(86) Multiple occurrence of carbonyl - A ring description contains more than one carbonyl.

(87) Single occurrence of carbonyl in more than one ring - More than one ring description contains one carbonyl.

(88) Multiple occurrence of carbonyl in more than one ring - More than one ring description containing more than one carbonyl.

(89) Single occurrence of exocyclic double bond - A ring description contains one exodouble bond.

(90) Multiple occurrence of exocyclic double bond - A ring description contains more than one exodouble bond.

(91) Single occurrence of exocyclic double bond in more than one ring - More than one ring description contains one exodouble bond.

(92) Multiple occurrence of exocyclic double bond in more than on ring - More than one ring description contains more than one exodouble bond.

(93) Single occurrence of any other heteroatom - Occurrence of any letter other than H, K, M, N, O, S, I, V, U, X, or Y.

(94) Multiple occurrence of any other heteroatom - Occurrence of any letter other than above more than once in the same description.

(95) Single occurrence of any other heteroatom in more than one ring - More than one ring description contains a letter other than those given above.

(96) Multiple occurrence of any other heteroatom in more than one ring- More than one ring description contains more than one letter other than those given above.

RING TYPES

On analysis of the WLN ring record, a ring type descpiption is set up which gives information on the size of each ring and the saturation/unsaturation value of that ring. The ring descriptor gives the atom types in each ring and this is used to determine whether hetero/carbo.

(97) Aromatic 6-membered ring - The presence of at least one 6-membered ring, fully unsaturated and no heteroatoms present in the ring description.

(98) Carbocyclic 5-membered ring - The presence of at least one 5-membered ring saturated or partially saturated and no heteroatoms present in the ring description.

(99) Carbocyclic 6-membered ring - The presence of at least one 6-membered ring, saturated or partially saturated, and no heteroatoms present in the ring description.

(100) Carbocyclic rings other than 5 and 6-membered - The presence of at least one ring (not 5 or 6-membered), saturated or partially saturated and no heteroatoms in the ring description

(101) Heterocyclic 5-membered ring - The presence of at least one 5-membered ring, saturated or unsaturated, and at least one heteroatom in the ring description.

(102) Heterocyclic 6-membered ring - The presence of at least one 6-membered ring, saturated or unsaturated, and at least one heteroatom in the ring description.

(103) Heterocyclic rings other than 5 and 6-membered - The presence of at least one ring (not 5 or 6-membered), saturated or unsaturated, and at least one heteroatom in the ring description.

HETEROATOM COUNT

Count of total number of heteroatoms of any type occurring in one ring.

(104) 1 heteroatom in one ring - Total of one heteroatom in one ring.

(105) 2 heteroatoms in one ring - Total of two heteroatoms in one ring.

(106) More than 2 heteroatoms in one ring - Total of three or more heteroatoms in one ring.

(107) 1 heteroatoms in more than one ring - Total of one heteroatom in more than one ring.

(108) 2 heteroatoms in more than one ring - Total of two heteroatoms in more than one ring.

(109) More than 2 heteroatoms in more than one ring - Total of three or more heteroatoms in more than one ring.

RING FUSIONS

A set of ring descriptions is set up for each ring system in the order in which they occur. These are compared to find the fusion types.

(110) 1 single heterocyclic ring - A heterocyclic ring unfused to any other ring.

(111) More than 1 single heterocyclic ring - More than one heterocyclic ring unfused to any other ring.

(112) 1 single carbocyclic ring - A carbocyclic ring unfused to any other ring.

(113) More than 1 single carbocyclic ring - More than one carbocyclic ring unfused to any other ring.

(114) 1 carbo/carbo fusion - A carbo ring (saturated or unsaturated) fused to a second carbo ring (saturated or unsaturated).

(115) More than 1 carbo/carbo fusion - More than 1 carbo ring attached to another carbo ring within the same ring system.

(116) 1 carbo/carbo fusion in more than 1 ring system - One carbo ring attached to a second carbo ring occurring in more than than one ring system.

(117) More than 1 carbo/carbo fusion in more than 1 ring system - More than 1 carbo/carbo fusion occurring in more than 1 ring system.

(118) 1 carbo/hetero fusion - A carbo ring (saturated or unsaturated) fused to a hetero ring.

(119) More than 1 carbo/hetero fusion - More than 1 carbo/hetero fusion occurring in the same ring system.

(120) 1 carbo/hetero fusion in more than 1 ring system - 1 carbo/hetero fusion in more than 1 ring system.

(121) More than 1 carbo/hetero fusion in more than 1 ring system - More than 1 carbo/hetero fusion occurring in more than 1 ring system.

(122) 1 hetero/hetero fusion - Two hetero rings fused to each other.

(123) More than 1 hetero/hetero fusion - More than one hetero/hetero fusion occurring in the same ring system.

(124) 1 hetero/hetero fusion in more than 1 ring system - 1 hetero/hetero fusion in more than 1 ring system.

(125) More than 1 hetero/hetero in more than 1 ring system - More than 1 hetero/hetero fusion occurring in more than 1 ring system.

RING LINKAGES

(126) Spiro ring indicator - Sequence locant-&locant in non-ring part of WLN.

DESCRIPTION OF THE PROGRAMMED CARCINOGENESIS KEYS
(Used Only in Carcinogenesis Model)

(127) True bridge indicator - WLN contains a ring notation with cited bridge locants.

(128) 1 multi-cyclic point - Within any ring signs sequence bna where n=1.

(129) More than 1 multi-cyclic point - Within any ring signs sequence bn where n > 1, or sequence bnn.

(130) Bilinkage - Two ring systems (including benzene) are linked together.

UNUSUAL CONDITIONS

(131) Chelate - WLN contains the character D. No other reliable fragments are set.

(132) Metallocene - Ring containing character zero, not within hyphens. Any other fragments set for metallocenes are not reliable.

(133) Inorganics - Notation begins with a space, but not ∧&&. No other fragments are set.

TOTAL RING FEATURES

Used to indicate the presence of ring features in the molecule.

(134) 1 ring system - Occurrence of one ring system (not benzene).

(135) 2 ring system - Occurrence of 2 ring systems (not benzene).

(136) More than 2 ring systems - Occurrence of more than 2 ring systems (not benzene).

(137) 1 benzene ring - Occurrence of one phenyl group.

(138) 2 benzene rings - Occurrence of 2 phenyl groups.

(139) More than 2 benzene rings - Occurrence of more than 2 phenyl groups.

(140) 1 carbocyclic ring - Occurrence of one individual fused or aromatic ring (excluding non-fused benzenes) in total molecule.

(141) 2 carbocyclic rings - Occurrence of two carbocyclic or aromatic rings (excluding non-fused benzenes) in total molecule.

(142) More than 2 carbocyclic rings - Occurrence of more than 2 carbocyclic or aromatic rings (excluding non-fused benzenes) in total molecule.

(143) 1 heterocyclic ring - Occurrence of one individual heterocyclic ring in total molecule.

(144) 2 heterocyclic rings - Occurrence of two heterocyclic rings in total molecule.

(145) More than 2 heterocyclics - Occurrence of more than 2 heterocyclic rings in total molecule.

SPECIAL COMPOUND TYPES

(146) Polypeptide - Notation begins with /. No other fragments are set.

(147) Polymer - Notation begins with /. No other fragments are set.

EXTENSIONS

(148) More than one 3-branch carbon atom - The character Y (but not ^Y) occurring more than once.

(149) Presence of suffix - A suffix beginning ^&& is present in the WLN.

ADDITIONAL METAL FRAGMENTS

These are found by locating the character sequence -AA- where AA is the metal WLN atomic symbol anywhere in the notation. Note: KA (potassium), WO (Tungsten), UR (uranium), VA (vanadium), and YT (yttrium) are not standard atomic symbols.

METAL	FRAGMENT	METAL	FRAGMENT	METAL	FRAGMENT
Ac	211	Hf	241	Pm	271
Al	212	He	242	Pa	272
Am	213	Ho	243	Ra	273
Sb	214	In	244	Ru	274
Ar	215	Ir	245	Re	275
As	216	Fe	246	Rh	276
At	217	Kr	247	Rb	277
Ba	218	La	248	Ru	278
Bk	219	Lr	249	Sm	279
Be	220	Pb	250	Sc	280
Bi	221	Li	251	Se	281
Cd	222	Lu	252	Si	282
Ca	223	Mg	253	Ag	283
Cf	224	Mn	254	Na	284
Ce	225	Md	255	Sr	285
Cs	226	Hg	256	Ta	286
Cr	227	Mo	257	Tc	287
Co	228	Nd	258	Te	288
Cu	229	Ne	259	Tb	289
Cm	230	Np	260	Tl	290
Dy	231	Ni	261	Th	291
Es	232	Nb	262	Tm	292
Er	233	No	263	Sn	293
Eu	234	Os	264	Ti	294
Fm	235	Pd	265	Wo	295
Fr	236	Pt	266	Ur	296
Gs	237	Pu	267	Va	297
Ga	238	Po	268	Ze	298
Ge	239	Ka	269	Yb	299
Hu	240	Pr	270	Yt	300
				Zn	301
				Zr	302

PROGRAMMED CARCINOGENESIS KEYS
(Used only in LD_{50}, Mutagenesis and Teratogenesis Models)

Note: These keys are generated as combinations of keys 1-309. They will often produce false positives, and so must be verified as correct. Keys 321, 322, 323, 324, 326, 329, and 332 are generated by searching for specific WLN character sequences.

(310) Aromatic amino - [Key 97 (aromatic 6-membered ring)] AND [Key 8 (3-branch nitrogen) OR Key 56 (substituent -NH-) OR Key 57 (> 1 substituent -NH-) OR Key 58 (substituent $-NH_2$) OR Key 59 (> 1 substituent $-NH_2$).]

(311) N-Nitroso, sulfonyl - [Key 161 (chain N-nitroso) AND Key 162 (chain sufonamide)] OR [Key 192 (substituent N-nitroso) AND Key 193 (substituent sulfonamide).]

(312) Organohalogen mustards - [Key 8 (3-branch nitrogen) OR Key 28 (chain -NH-) OR Key 29 (> 1 chain -NH-) OR Key 56 (substituent -NH-) OR Key 57 (> 1 substituent-NH-)] AND [Key 19 (chain ethyl/ethylene) AND Key 22 (chain halogen)].

(313) Organohalogen mustards - [Key 10 (sulfur atom) OR Key 11 (> 1 sulfur)] AND [Key 19 (chain ethyl/ethylene)] AND [Key 22 (chain halogen).]

(314) α Haloether - [Key 6 (3-branch carbon) OR Key 148 (> 1 3-branch carbon) OR Key 7 (4-branch carbon) OR Key 17 (chain methyl/methylene) OR Key 18 (> 1 chain methyl/methylene) OR Key 20 (chain alkyl, 3-9 carbons) OR Key 21 (chain alkyl, 10 or more carbons) OR Key 19 (chain ethyl/ethylene)] AND [Key 22 (chain halogen)] AND [Key 35 (chain oxygen) OR Key 36 (> 1 chain oxygen).]

(315) Haloalkane - [Key 6 (3-branch carbon) OR Key 148 (> 1 3-branch carbon) OR Key 7 (4-branch carbon) OR Key 17 (chain methyl/methylene) OR Key 18 (> 1 chain methyl/methylene) OR Key 19 (chain ethyl/ethylene) OR Key 20 (chain alkyl, 3-9 carbons) OR Key 21 (chain alkyl, 10 or more carbons)] AND [Key 22 (chain halogen).]

(Note: Key 343 is α -β dihaloalkanes, Key 344 is geminal dihaloalkanes and Key 345 is trihaloalkanes).

(316) β -Haloether - [Key 6 OR 148 (3-branch carbon) OR Key 7 (4-branch carbon) OR Key 17 or 18 (chain methyl/methylene) OR key 19 or 20 (chain alkyl)] AND [Key 22 (chain halogen)] AND [Key 35 or 36 (chain oxygen).]

(317) α -Haloalkene - [Key 6 or 148 (3-branch carbon) OR Key 17 or 18 (chain methyl/methylene) OR Key 45 or 46 (substituent methyl/methylene)] AND [Key 14 or 15 (double bond)] AND [Key 22 (chain halogen).]

(Note:Key 346 is α -haloalkene)

(318) Halogenated aromatic - [Key 50 (substituent halogen)] AND [Key 97 (aromatic 6-membered ring).]

(319) Alkyl sulfate - [Key 4 (dioxo)] AND [Key 10 or 11 (sulfur).]

(320) Sultone - [Key 4 (dioxo)] AND [Key 73 (oxygen heteroatom)] AND [Key 81 (sulfur heteroatom)] AND [Key 105 or 108 (2 heteroatoms)].

(321) Epoxide - WLN character sequence T3OTJ or T4OTJ.

(322) Aziridene - WLN character sequence T3NTJ, T3MTJ, T3KTJ, T4NTJ, T4MTJ, or T4KTJ.

(323) Episulfide - WLN character sequence T3STJ or T4STJ.

(324) β -Lactones and lactams - WLN character sequence T4OVTJ, T4NVTJ, T4MVTJ.

(325) (not used)

(326) 5-membered ring α,β unsaturated lactone - WLN character sequence T5OV CUTJ or T5OV CUJ.

(327) 5-membered ring anhydrides - [Key 73 (oxygen heteroatom)] AND [Key 86 (> 1 carbonyl hetero group)] AND [Key 101 (5-membered heterocyclic ring)] AND [Key 104 or 107 (1 heteroatom-carbonyls are not counted)].

(328) α,β unsaturated carbonate - [Key 74 (> 1 oxygen heteroatom)] AND [Key 85 or 86 (carbonyl hetero group)] AND [Key 105 or 108 (2 heteroatoms)] AND [Key 101 (5-membered heterocyclic ring)]

(329) 6-membered ring α,β unsaturated lactone - WLN character sequence T6OV CUTJ.

(330) Fused aromatic α,β unsaturated lactone - [Key 73 (oxygen heteroatom)] AND [Key 97 (6-membered aromatic ring)] AND [Key 118 (1 carbo/hetero fusion)] AND [Key 104 or 107 (single heteroatom)].

(331) Fused polynuclear aromatic - [Key 97 (6-membered aromatic ring)] AND [Key 115 (> 1 carbo/carbo fusion) OR Key 116 (1 carbo/carbo fusion in 1 ring system) OR Key 117 (> 1 carbo/carbo fusion in > 1 ring system)].

(332) Aryldialkatriazene - WLN character sequence (N)U(N)(N), (N)(N)U(N),or (N)_(N)U(N).

(333) Purine analog - [Key 5 (terminal oxygen)] AND [Key 80 (> 1 N in > 1 ring) AND Key 101 (5-membered heterocyclic ring) AND Key 102 (6-membered heterocyclic ring) AND Key 108 (2 heteroatoms in > 1 ring) AND Key 122 (1 hetero/hetero fusion)].

(334) 6-membered heterocyclic ring with 2 nitrogens - [Key 78 (> 1 N in one ring)] AND [Key 102 (6-membered heterocyclic ring)] AND [Key 105 (2 heteroatoms in 1 ring)] AND [Key 110 (single heterocyclic ring) OR Key 111 (more than 1 single heterocyclic ring)].

(335-336) Used only for carcinogenicity models

(335) Any one or more of the following keys:
100, 207, 223, 285, 314, 330, 332.

(336) Any one or more of the following keys:
21, 42, 94, 176, 281, 309.

(337) Acylated aromatic amide - [Key 97 (aromatic 6-membered ring)] AND [Key 182 (substituent secondary amide) OR Key 183 (substituent tertiary amide)].

(338) Aromatic hydroxylamino - [Key 97 (aromatic 6-membered ring) AND Key 199 (substituent hydroxylamine)].

(339) Aromatic nitroso - [Key 97 (aromatic 6-membered ring) AND Key 191 (substituent C-nitroso)].

(340) Aromatic azo - [Key 97 (aromatic 6-membered ring) AND Key 190 (substituent azo)].

(341) Aromatic nitro - [Key 97 (aromatic 6-membered ring) AND Key 3 (nitro group)].

(342) N-nitroso amide - [Key 161 (chain N-nitroso) AND Key 151 (chain secondary amide) OR Key 152 (chain tertiary amide)] OR [Key 192 (substituent N-nitroso) AND (Key 182 (substituent secondary amide) OR Key 183 (substituent tertiary amide)].

(343) α,β - dihaloalkane - [Key 6 (single 3-branch carbon) OR Key 148 (> 1 3-branch carbon) OR Key 7 (4-branch carbon) OR Key 17 (1 methyl group) OR Key 18 (> 1 methyl group)] AND [Key 22 (chain halogen)].

(344) Geminal-dihaloalkane - [Key 6 (single 3- branch carbon) OR Key 148 (> 1 3-branch carbon) OR Key 7 (4-branch carbon) OR Key 17 (1 methyl group) OR Key 18 (> 1 methyl group)] AND [Key 22 (chain halogen)].

(345) Trihaloalkane - [Key 6 (single 3-branch carbon) OR Key 7 (4-branch carbon) OR Key 148 (> 1 3-branch carbon)] AND [Key 22 (chain halogen)].

(346) β - Haloalkene - [Key 6 (3-branch carbon) OR Key 148 (> 1 3-branch carbon) OR Key 7 (4-branch carbon) OR Key 19 (chain ethyl) AND Key 14 (double bond) OR Key 15(> 1 double bond)] AND [Key 22 (chain halogen)].

(Note: Key 317 is α -haloalkene.

(347) Unfused poly-chlorinated alicyclic - [Key 52 (> 1 substituent chlorine)] AND [Key 112 (1 carbocyclic ring) OR Key 113 (> 1 unfused carbocyclic ring)].

(348) Fused poly-chlorinated alicyclic - [Key 52 (> 1 substituent chlorine)] AND [Key 114 (1 carbo/carbo fusion) OR Key 115 (> 1 carbo/carbo fusion) OR Key 116 (1 carbo/carbo fusion in > 1 ring system) OR Key 117 (> 1 carbo/carbo fusion in > 1 ring system) OR Key 118 (1 carbo/hetero fusion) OR Key 119 (> 1 carbo/hetero fusion) OR Key 120 (a carbo/hetero fusion in > 1 ring system) OR Key 121 (> 1 carbo/hetero fusion in > 1 ring system)].

(349) Polychlorinated biphenyl - [Key 52 (>1 substituent chlorine)] AND [Key 180 (biphenyl).

(350) Hydrazo/hydrazine - [Key 28 (chain -NH-) OR Key 56 (substituent -NH-) AND Key 30 (chain $-NH_2$) OR Key 31 (> 1 chain $-NH_2$)] OR [Key 29 (> 1 chain -NH-) OR Key 57 (> 1 substituent -NH-)].

DESCRIPTION OF THE PROGRAMMED CARCINOGENESIS KEYS
(Used Only in Carcinogenesis Model)

(310) Aromatic amino - [Key 97 (aromatic 6-membered ring)] AND [Key 8 (3-branch nitrogen) OR Key 56 (substituent N-H bond) OR Key 58 (amino group) OR Key 59 (> 2 amino group)].

(311) N-nitroso, sulfonyl - [Key 161 (chain N-nitroso)] AND [Key 162 (chain sulfonamide)].

(312-314) Organohalogen mustards

(312) [Key 8 (3-branch nitrogen) OR Key 28 (chain N-H bond) OR Key 56 (substituent N-H bond)] AND [Key 19 (ethyl or ethylene group)] AND [Key 22 (generic halogen)].

(313) [Key 10 (sulfur atom)] AND [Key 19 (ethyl or ethylene group)] AND [Key 22 (generic halogen)].

(314) [Key 313] AND [Key 36 (> 1 chain oxygen)].

(315) Halo alkanes - [Key 6 (3-branch carbon) OR Key 7 (4-branch carbon) OR Key 17 (methyl or methylene group) OR Key 19 (ethyl or ethylene group) OR Key 20 (alkyl chain $(CH_2)_n$, n=3-9) OR Key 148 (> 1 3-branch carbon)] AND [Key 22 (generic halogen)].

(316) Haloethers - [Key 6 (3-branch carbon) OR Key 7 (4-branch carbon) OR Key 17 (methyl or methylene group) OR Key 19 (ethyl or ethylene group) OR Key 148 (> 1 3-branch carbon)] AND [Key 35 (1 chain oxygen)] AND [Key 22 (generic halogen)].

(317) Haloalkenes - [Key 6 (3-branch carbon) OR Key 17 (methyl/methlylene group)] AND [Key 14 (carbon double bond, not C=OH, C=N, C=S] AND [Key 22 (generic halogen)].

(318) Halogenated aromatics - [Key 50 (substituent halogen)] AND [Key 97 (aromatic 6-membered ring)] AND [Key 180 (biphenyl) OR Keys 114 thru 125 (any one or more) various types and amount of ring fusions in overall compound].

(319) Alkyl sulfates - [Key 4 (dioxo group) OR Key 36 (> 1 oxygen)] AND [Key 10 (sulfur)].

(320) Sulfones - [Key 4 (dioxo group)] AND [Key 73 (oxygen as ring heteroatom)] AND [Key 81 (sulfur as ring heteroatom)] AND [Keys 104 or 107 (1 heteroatom in 1 or > 1 ring)].

(321) Epoxides - [Key 103 (heterocyclic ring, not 5- or 6-membered)] AND [Key 73 (oxygen as ring heteroatom)] AND [Keys 204 or 107 (1 heteroatom in 1 or > 1 ring)].

(322) Aziridines - [Key 103 (heterocyclic ring, not 5- or 6-membered)] AND [Key 77 (nitrogen as ring heteroatom)] AND [Keys 104 or 107 (1 heteroatom in 1 or > 1 ring)].

(323) Sulfides - [Key 103 (heterocylic ring, not 5- or 6-membered)] AND [Key 81 (sulfur as ring heteroatom)] AND [Keys 104 or 107].

(324) β - Lactones - [Key 321 (epoxides)] AND [Key 85 (carbonyl in ring)].

(325) β - Lactam - [Key 322 (Azirides)] AND [Key 85 (carbonyl hetero group)].

(326) β -Unsaturated lactones - [Key 101 (heterocyclic 5-membered ring)] AND [Key 73 (oxygen as ring heteroatom)] AND [Key 85 (carbonyl in ring)] AND [Keys 104 or 107 (a heteroatom in 1 or > 1 ring)].

(327) Anhydrides - [Key 101 (heterocyclic 5-membered ring)] AND [Key 73 (oxygen as ring heteroatom)] AND [Key 86 (> 1 carbonyl in ring)] AND [Keys 104 or 107 (1 heteroatom in 1 or > 1 ring)].

(328) α - β Unsaturated carbonates - [Key 101 (heterocyclic 5-membered ring)] AND [Key 24 (> 1 oxygen as ring heteroatom)] AND [Key 85 (carbonyl in ring)] AND [Keys 105 or 108 (2 heteroatoms in one or > 1 ring)].

(329) α - β Unsaturated lactones - [Key 102 (heterocyclic 6- membered ring)] AND [Key 73 (oxygen as ring heteroatom)] AND [Key 85 (carbonyl in ring)] AND [Keys 104 or 107 (1 heteroatom in 1 or > 1 ring)].

(330) Fused aromatic α - β unsatureated lactones - [Key 329] AND [Key 97 (aromatic 6- membered ring)] AND [Key 118 (1 carbo/hetero fusion)].

(331) Fused polynuclear aromatics - [Keys 114 thru 125 (any one)] AND [Key 97 (aromatic 6-membered ring)].

(332) Aryldialkatriazenes - [Key 97 (aromatic 6-membered ring)] AND [Key 204 (substituent phenylazo)] AND [Key 8 (3-branch nitrogen)].

(333) Purine analog - [Key 80 (> 1 nitrogen as heteroatom in 1 ring)] AND [Key 101 (heterocyclic 5- membered ring)] AND [Key 102 (heterocyclic 6-membered ring)] AND [Key 108 (2 heteroatoms in > 1 ring)] AND [Key 122 (1 hetero/hetero fusion)].

(334) Pyrimidine analogs - [Key 78 (> 1 nitrogen as heteroatom in 1 ring)] AND [Key 102 (heterocyclic 6-membered ring)] AND [Key 105 (2 heteroatoms in 1 ring)].

(335) Any one or more of the following keys: 100, 207, 223, 285, 314, 330, 332.

(336) Any one or more of the following keys: 21, 42, 94, 176, 281, 309.

APPENDIX E
EXAMPLE OF GENERATING AN LD_{50} ESTIMATE

1. Compound: Malononitrile, o-chlorobenzylidene

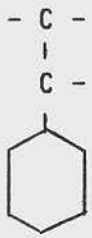
2. RTECS No.: 003675000

3. CAS No.: 2698411

4. Molecular Formula: $C_{10}H_5ClN_2$

5. Molecular Weight: 188.6

6. Molecular Structure:



7. WLN: NCYCN&U1R BG

8. Keys Generated:
6^{L}, 8^{CM}, 14^{LM}, 17^{LC}, 34^{L}, 50^{LT}, 51^{LC}, 97^{-}, 137^{CM}, 179^{-}

9. Table of key coefficient values (LD_{50} designated keys only):

KEY NO.	COEFFICIENT VALUE
6	.096
14	.141
17	.089
34	.278
50	.212
51	-.098
	=.718

10. Obtain log value of molecular weight: Log of 188.6 = 2.276

11. Obtain log of molecular weight and then multiply by constant:

(2.276)(.681) = 1.55

12. Determine the natural log of the reciprocal of the concentration value:

log 1/c = coefficient values of keys + regression constant
log 1/c = .718 + 1.55 + .552
log 1/c = 2.820

13. Determine antilog of log 1/c: antilog of 2.820 = 661

14. Obtain LD_{50} endpoint estimate as follows:

$$LD_{50}\ (mg/kg) = \frac{(mol\ wt)\ (1000)}{antilog\ log\ 1/c} = \frac{(188.6)\ (1000)}{661} = 285\ mg/kg$$

Note: Based on biological data from the literature the LD_{50} for this compound is thought to be 178 mg/kg.

APPENDIX H
EXAMPLE OF GENERATING AN ESTIMATE OF MUTAGENICITY

1. Compound: Malononitrile, o-chorobenzylidene

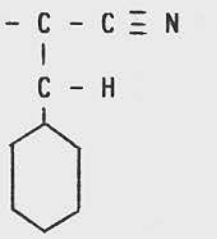
2. RTECS No.: 003675000

3. CAS No.: 2698411

4. Molecular Formula: $C_{10}H_5ClN_2$

5. Molecular Weight: 188.6

6. Molecular Structure:



7. WLN: NCYCN&U1R BG

8. Keys Generated:
6^{L}, 8^{CM}, 14^{LM}, 17^{LC}, 34^{L}, 50^{LT}, 51^{LC}, 97^{-}, 137^{CM}, 179^{-}

9. Table of key coefficient values (only Mut designated keys):

KEY NO.	COEFFICIENT VALUES	
	POS	NEG
8	1.668	-0.489
14	3.746	-1.209
137	2.134	3.771
regession constants	-5.078	-3.183

10. Probability equation expressed in exponential terms:

$$\text{Probability of mutation} = \frac{e^{expm+}}{e^{expm+} + e^{expm-}}$$

11. Conversion of exponential values to natural logs:

2.47 = 11.9
1.308 = 3.7

12. Probability value determined as follows:

$$\frac{11.9}{11.9 + 3.7} = .76 \text{ probability of mutagenicity}$$

APPENDIX J
EXAMPLE OF GENERATING AN ESTIMATE OF CARCINOGENICITY

1. Compound: Malononitrile, o-chlorobenzylidene

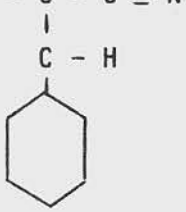
2. RTECS No.: 003675000

3. CAS No.: 2698411

4. Molecular Formula: $C_{10}H_5ClN_2$

5. Molecular Weight: 188.6

6. Molecular Structure:



7. WLN: NCYCN&U1R BG

8. Keys Generated:

6^{L}, 8^{CM}, 14^{LM}, 17^{LC}, 34^{L}, 50^{LT}, 51^{LC}, 97^{-}, 137^{CM}, 179^{-}

9. Table of key coefficient values (only Car designated keys):

KEY NO.	COEFFICIENT VALUES	
	DEFINITE	INDEFINITE
Mol. Wt.	.02	.01
17	1.98	3.66
51	-1.17	4.24
137	1.32	5.83
regression constants	-6.55	-7.42
	- .21	5.69

10. Probability equation expressed in exponential terms:

$$\text{Probability of carcinogen} = \frac{e^{expc+}}{e^{expc+} + e^{expc-}}$$

11. Conversion of exponential values to natural logs:
    -.21 = 0.811
    5.69 = 195.89

12. Probability value determined as follows:

$$\frac{0.811}{0.811 + 195.89} = .005 \text{ probability of carcinogenicity}$$

APPENDIX L
EXAMPLE OF GENERATING AN ESTIMATE OF TERATOGENICITY

1. Compound: Diethyl phthalate

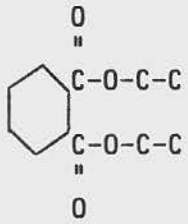
2. RTECS No.: TI1050000

3. CAS No.: 84662

4. Molecular Formula: $C_{12}H_{14}O_4$

5. Molecular Weight: 222.26

6. Molecular Structure:



7. WLN: 2OVR BVO2

8. Keys Generated:

   19^{TM}, 72^{T}, 97^{-}, 137^{M}

9. Table of key coefficient values (only Ter designated keys):

KEY NO.	COEFFICIENT VALUES	
	DEFINITE	INDEFINITE
19	2.362	3.507
72	0.628	-2.590
regression constants	-3.667	-4.406
	-0.677	-3.489

10. Probability equation expressed in exponential terms:

$$\text{Probability of teratogen} = \frac{e^{expt+}}{e^{expt+} = e^{expt-}}$$

11. Conversion of exponential values to natural logs:
    -0.677 = 0.508
    -3.489 = 0.031

12. Probability value determined as follows:

$$\frac{0.508}{0.508 + 0.031} = .943 \text{ probability of teratogenicity}$$

Errata page for the NIOSH technical report "The Development and Application of Algorithms for Generating Estimates of Toxicity for the NOHS Data Base"

p. viii - Abstract, 2nd paragraph, last line, should read "...little or no toxicity data have been reported...".

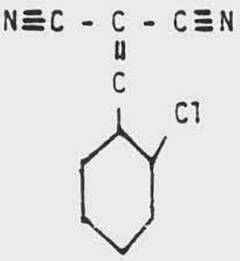
p. 4 - B. Modeling the Algorithms, 2nd paragraph, line 11, is more accurate if it reads "...molecular structure..." than "...molecular formula...".

p. 7 - last paragraph, line 8, should read 350 keys rather than 359 keys. This change should be noted throughout the report wherever 359 keys appears.

p. 8 - C. Statistical Methodologies, line 4, should read "...analysis was used based on...".

p. 35 - first paragraph, line 3, comma should follow after Figure 7.

pp. 93, 94, and 95 - Appendices E, H, and J, item 6, the correct molecular structure is as follows:



Note: hydrogen atoms not shown

p. 95 - Appendix J - Example of Generating an Estimate of Carcinogenicity, items 9, 11, and 12 should be changed to include key 8. These items should then read as follows:

9. Table of key coefficient values (car designated keys):

Key No.	Coefficient Value	
	Definite	Indefinite
8	4.19	-0.63
17	1.98	3.66
51	-1.17	4.24
137	1.32	5.83
Molecular Weight	0.02	0.01
Regression Constants	-6.55	-7.42
	-0.29	5.69

11. Conversion of exponential values to natural logarithms:

-0.29 = 0.75
5.69 = 295.90

12. Probability value determined as follows:

$$\frac{0.75}{0.75 + 295.90} = 0.002 \text{ probability of carcinogenicit}$$