Indian Journal of Animal Sciences **83** (8): 838–843, August 2013/Article

Animal experiments data: Structural analysis and design a model for determining lethal drug dosage

A BABANLI[1] and A GUNES[2]

Istanbul Aydin University, Besyol Mah. Kucukcekmece, Istanbul, Turkey

Received: 1 August 2012; Accepted: 1 March 2013

ABSTRACT

The ACD-IDEA database, originally developed by the authors, stores main results of scientific publications on animal experiments realized around the world during the last 100 years. In this paper a common statistical analysis of this database is presented. On the basis of the knowledge extracted from this database, a rule-based model is developed that is able to estimate lethal drug dosages for a given groups of compounds and animals. A decision tree is designed and a set of association rules with exceptions is created. This model can be used for determining lethal dosage values for drugs of the category "General anesthetic compounds" and for animals of the group "rabbit, cat, dog, guinea pig". This model can also be used in a diverse set of problems of pharmacological research and education.

Key words: Animal experiment data, Data analysis, Decision tree, Rule-based model

As the volume of data increases, the proportion of this data that people understand decreases. Lying hidden in all this data is information that is rarely made explicit or taken advantage of (Witten *et al.* 2005). One way to extract useful information from complex data sets is to look for patterns in data. A scientist's job is to make sense of data, to discover the patterns that govern how the world works and encapsulate them in theories that can be used for prediction. Patterns generally express structural description of data sets. The acquisition of descriptions from examples can be realized using machine learning methods. The kind of structural descriptions found can be used for explanation, understanding and prediction.

There is a tremendous amount of data on pharmaceutical research results readily available in scientific publications; the rate of which increasing year-on-year (Boyer 2009). Two editions of "Drug Dosage in Laboratory Animals: A Handbook" (Barnes *et al.* 1973, Borchard *et al.* 1990), contain a wide range of existing data from relevant laboratory experiments on animals. Based on these data the database ACD-IDEA was developed (Gunes *et al.* 2009). The database is a great repository of raw data that can be used for extraction of a variety of useful information for pharmacological education and research, some of which were considered by Babanli *et al.* (2011).

In this paper we present the results of statistical investigations of some common aspects of drug-animal-dosage-action relationships. Furthermore; we describe a rule-based model that can be used to estimate lethal toxicity doses for some group of animals and drugs.

Present address: [1,2]Researcher (ahmetbabanli@aydin.edu.tr, aligunes@aydin.edu.tr).

MATERIAL AND METHODS

Data analysis: The database ACD-IDEA contains 4,273 records, each corresponding to a single animal experiment result. Records of the database have the following structure:

Drug, Animal, Administration Route, Dose (lower and upper bounds), Action, Reference (publication from which the data were taken). We'll use the first five attributes in our investigations denoting them as Drug, Animal, Route and Dose Lethal; where the attribute Action is renamed to be Lethal for clarity.

In the database the greatest part of records have equal values for lower and upper bounds of Dose. For this reason lower bound is used in its calculations. Codes were used rather than unit names in many cases (see Appendices). A simple statistical analysis of the database gives numerous useful information on the inner structure of datasets. Main results of such investigation are given in Tables 1–4.

Table 1 has total number of records (instances in data mining terminology) for every category of compounds (drugs). In the database all compounds are stored in the tree-like scheme with four levels (Babanli 2011). The data of the Table correspond to the highest level (List A).

Table 2 has the total number of instances for animal actions. A two-level hierarchy was designed for actions. In

Table 1. The number of instances for compounds

C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C14	C15	C16
3412	156	17	86	74	13	23	-	10	21	-	218	25	194	17

Table 2. The number of instances for actions

R01	R02	R03	R04	R05	R06	R07	R08	R09	R10
1884	294	84	293	68	11	32	27	15	4
R11	R12	R13	R14	R15	R16	R17	R18	R19	
5	-	73	1	25	6	1	1334	116	

Table 3. The number of instances for animals

B	C	D	M	N	P	R
571	568	662	1075	179	280	938

Table 4. The number of instances for routes

IM	IP	IV	PO	SC
253	855	1456	834	875

List B the first level appears. Table 3 contains the number of instances for animals (List C). Table 4 contains the number of instances for administration routes (List D). From the tables one might conclude:

– In animal experiments drugs of the category "Compounds acting on the nervous system" were used in numerous cases: 3,412 times from a total of 4,273. Drugs of type "Antidotes" were only used in 7 experiments.

A. Compounds of the first level

Code	Compounds
C01	Acting on nervous system
C02	Acting on cardiovascular system
C03	Affecting blood and blood formation
C04	Acting on respiratory tract
C05	Acting on gastrointestinal tract
C06	Acting on urinary tract
C07	Acting on reproductive tract
C08	Affecting liver and bile production
C09	Acting on hormonal system
C10	Acting on immune system
C11	Dermatological
C12	Autacoids and related
C13	Chemotherapeutic
C14	Antidotes
C15	Having no therapeutic categories
C16	Miscellaneous

B. Actions of the first level

Code	Action
R01	Action on the central nervous system
R02	Action on the autonomic nervous system
R03	Action on skeletal muscle function
R04	Action on the cardiovascular system
R05	No steroidal analgesic, anti-pyretic and anti-inflammatory action
R06	Action on blood homodynamic
R07	Action on the kidneys
R08	Action on the gastrointestinal tract
R09	Action effective on the respiratory tract
R10	Action effective on the reproductive system
R11	Action effective on smooth muscle
R12	Action effective on the immune system
R13	Action effective on histamine, serotonin and bradykinin metabolism
R14	Action effective on liver function
R15	Action effective on metabolic function
R16	Action effective on cell proliferation and genes
R17	Antidotes
R18	Toxic effects
R19	Miscellaneous

C. Animals

Code	Animal
B	Rabbit
C	Cat
D	Dog
M	Mouse
N	Monkey
P	Guinea pig
R	Rat

D. Administration routes

Code	Route
IM	Intramuscular
IP	Intraperitoneal
IV	Intravenous
PO	Per oral
SC	Subcutaneous

E. Drugs of the "General anesthetic compounds"

Code	Drug
01	Chloral hydrate
02	Alpha-choralose
03	Hexobarbital
04	Paraldehyde
05	Phencyclidine
06	Probarbital
07	Secobarbital
08	Thiopental
09	Tribromoethanol
10	Uretan

"Dermatological compounds and compounds affecting liver and bile production" were not used at all.

– The greatest part of experiments (3,218) resulted in animal actions of two types: "Actions on central nervous system" and "Toxicity"- 1,884 and 1,334 respectively. Only one experiment was observed with "Actions effective on liver functions" and no experiment with "Action effective on immune system".
– Small animals (mouse and rat) were used more frequently: 1,075 and 938 respectively. The number of experiments on monkeys was small (179).
– The administration route IV was used more frequently- in 1,456 experiments and IM- less frequently- 253 times. Other routes- IP, PO, and SC were used approximately equally- about 850 times.

Creating a training dataset

The experimental data of the database was used for designing a model that can allow determining approximations to lethal toxicity doses that are very important in animal experiments. Several computer systems were designed to determine drug dosage in disease's treatment (Watton *et al.* 1999, Durieux *et al.* 2010). The object of our investigations is deciding drug dosage in animal experiments and is based on the results of a great amount of completed experiments performed around the world. Because of complexity of relationships between animal types, administration routes, drugs and doses the data seem have very higher heterogeneity in inner structures and so in these situations taking routine decisions is very difficult. For this reason after detailed analysis of data it was decided to look for decisions for a restricted group of drugs and animals. We have selected drugs of category "General anesthetic compounds" (List E) and the group of "non-small" animals (rabbit, cat, dog and guinea pig) for which the database contains 138 records. This set is used as input dataset for our further investigations.

Data cleaning is a necessary procedure for successful data mining. With a large data collection one should sample a few instances (called a training dataset) and examine them using some method. The received results are to be applied to a whole dataset and corrected if necessary.

Table 5. The training dataset

Drug	Animal	Route	Dose	Lethal
01	B	PO	30	n
01	B	SC	1000	y
01	C	IV	300	n
01	C	PO	440	y
01	D	IV	40	n
02	B	IV	120	n
02	B	SC	80	y
02	C	PO	600	y
02	D	IV	100	n
03	B	IP	225	y
03	B	IV	80	y
03	C	IP	40	n
03	C	PO	400	y
03	D	IV	30	n
07	B	IV	45	y
07	B	SC	50	n
07	C	IP	75	y
07	C	IV	54	n
07	D	PO	40	n
07	D	PO	90	y

Instances (20) were selected to create the training dataset (Table 5). In many cases for heterogeneous data machine learning methods gave non tractable results (Witten *et al.* 2005). For this reason selection included those instances to the dataset that make its inner structures as homogeneous as possible.

In the training dataset the instances take values from the following domains: Drug- from [01, 02, 03, 07], Animal – from [B, C, D], Route – from [IP, IV, PO, SC]. For lethal binary values are used: n – for *no,* meaning non-lethal case, and y – for *yes,* meaning lethal case (because of this reason it was renamed Action). Lethal is used as class attribute in machine learning methods we have applied. Drug, animal and lethal are all nominal attributes. Many machine learning methods require attributes of nominal types. But the attribute dose has numeric values. These were converted to nominal form. In the dataset Dose takes its numeric values from the range [30–1,000]. The following simple discretization method was used for conversion. First, all the values were sorted in ascending presenting repeated values once. This produces the sequence

					64.5			95				350			
30	40	45	50	54	75	80	90	100	120	225	300	400	440	600	1000
2n	3n	1y	1n	1n	1y	2y	1y	1n	1n	1y	1n	1y	1y	1y	1y

Once sorted, the number of instances was determined for each value of classes y and n. These numbers are shown on the next line where, for example, 2n means 2 instances of class n (for dose 30). Placing breakpoints in the sequence completes the discretization; place breakpoints wherever

class changes. To decrease the number of breakpoints it is convenient to place them so that on either side of a breakpoint lay instances of a majority class. Placing the breakpoint between 54 and 75 left side was having majority n-classes (7n and 1y) and right side with 4y. Another two possible breakpoints will be between 90 and 100 and between 300 and 400. As a result 4 categories were received. Breakpoint values can be calculated choosing half ways between left and right values. So, breakpoint values are 64.5, 95, and 350.

Now 4 nominal values can be determined for the Dose as

D1: Dose < 64.5

D2: Dose is in interval [64.5 – 95]

D3: Dose is in interval [95 – 350]

D4: Dose > 350.

From now on these values will replace the numeric values of the Dose column in Table 5.

Decision tree: The aim was to look whether there are any structural shapes (patterns) expressing relationships between the four attributes and the class attribute. If so, one can create a set of rules enabling to define borders between classes y and n. One approach towards solution to this problem is to design a decision tree (Russel *et al.* 2003, Witten *et al.* 2005). Input to the method would be a set of independent instances. Records of the database can be considered as independent instances, so we may use them.

The idea of the method is as follows: Take the set of instances as input, split them into pieces, send part of it down each branch and from there right on down to the leaves of the subtrees involved. The split is accomplished using attributes one by one.

Branches of the division correspond to values of the attribute. Every step is to give the best discrimination of classes y and n. So, it is necessary to determine the best attribute at any split. For this the error rate of the split rule is calculated.

An error rate is a relation of type *a*/*b*, where *a* is the number of erroneously classified instances and *b* is the total number of instances. Example for Drug is shown in Table 6.

The attribute Drug has four nominal values: 01, 02, 03, and 07. Consider the case for 01. There are 5 instances with this value (Table 5) 3 of which belong to class n and 2 to class y. So n is a majority class. The expression 01 -> n is a rule to find instances of class n. Applying it to 5 instances gives 3 true (of class n) and 2 wrong (of class y) instances, so we say that error rate of the rule 01 -> n is 2/5. Error rates for another rules are calculated similarly (when the numbers

Table 6. Error rates evaluated for drug

Split rule	Error	Total error
01 -> n	2/5	9/20
02 ->y	2/4	
03 -> y	2/5	
07 -> y	3/6	

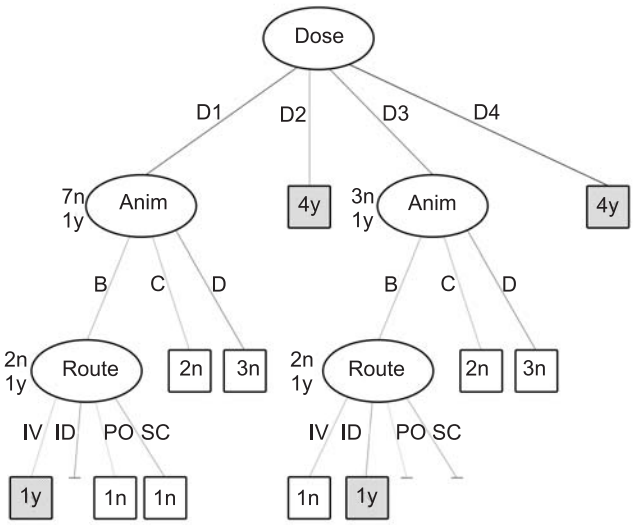

Fig. 1. Decision tree.

of y and n instances are equal, either of classes can be taken as majority). The total error for the Drug is defined as 9/20, where 9 is the number of errors and 20 is the number of all instances in the training dataset. Calculation of error rates for other attributes gives: Animal – 6/20, Route – 6/20 and Dose – 2/20. The Dose has the smallest error rate and it is to be taken as the best attribute to split the dataset at the first step.

In Fig. 1 the whole decision tree is given. At the first level the dataset is divided into 4 branches according to the values of Dose. Branches have 8, 4, 4, 4 instances respectively. The instances of the second and fourth branches are of the same classes and so they have no need to be classified further. In the first and third branches we have mixed instances and it is necessary to split them further by using the other attributes. The first branch contains 8 instances – 7 of n and 1 of y. We can easily classify them without calculating error rates. Application of the attribute Animal gives 3 new branches at the next level. One of them has the mixed subset – 2n and 1y. We select Route and split it further yielding the left branch to be classified completely. The third branch is constructed following similar steps.

Creating the rules

The decision tree of the Fig. 1 can be used for estimating lethal doses in animal experiments for the given set of drugs, animals and routes. But a rule set created on the base of it would be more understandable and applicable in practice.

Visual observation of the decision tree allows inferring:

- 18 of 20 instances are classified correctly at the first level.
- Branches D2 and D4 have instances of class y without errors.
- Branches D1 and D3 have the majority class n each with one error.
- At the second level the search spaces are decreased but

errors exist yet.

– All instances are classified correctly using 3 of 4 attributes. Creating the following two rules based on the first level is straightforward:
if (Dose in [D1, D3]) then Lethal= no.
if (Dose in [D2, D4]) then Lethal= yes.

Here, for example, the 1st rule means: if Dose is D1 or D3 then class value is no.

For the training dataset the 2nd rule covers all the instances whereas the 1st has two errors. To get a more general picture, which can be considered as validation of the results, we have applied them to the whole dataset (61 instances). Calculating error rates for Drug values gives: 01 – 1/21, 02 – 3/9, 03 – 3/15, 07 – 4/16.

The best result is for drug 01: 1 error of 21 instances and the worst is for drug 02: here one- third is classified wrongly. These require creating (modifying) the set of rules having exceptions that will be able to decrease the number of errors. This stage can be viewed as an approach to generalization. After adding exceptions the following 4 resulting rules were formulated that cover 60 from a set of 61 instances.

R1: *if (Dose < 64.5) then Lethal = no*
except if (Drug = 07 and Dose > 45 and Anim in [BIP,BIV,CIV,CPO])
then Lethal = yes.

R2: *if (Dose = between(64.5 – 95)) then Lethal = yes*
except if (Drug = 02 and Dose = 75 and Anim = CIV)
then Lethal = no.

R3: *if (Dose = between(95 – 350)) then Lethal = no*
except if ((Drug = 03 and ((Dose > 100 and Anim in [CIV,DIV])
or (Dose > 225 and Anim = BIP))
or
(Drug = 02 and Dose > 150 and Anim = CIP))
then Lethal = yes.

R4: *if (Drug > 350) then Lethal = yes*
except if (Drug = 01 and Dose > 500 and Anim = DPO) Lethal = no.

Explanations:

– Drug values are used in the numeric form for understandability.
– Expression like Dose = between (…) means that Dose takes values from an interval (…).
– Value like BIP denotes values for two attributes: for Animal and for Route respectively.
– Anim means Animal.
– In applications in the case of expression like Dose = m, where m is a numeric value, some range like m±e would be taken instead of an exact value (e can be taken by a researcher according to dose values and experiment features). For example, it is more realistic to use Dose = 75±5 rather than Dose = 75. This is also recommended for other comparison operations like > and <.

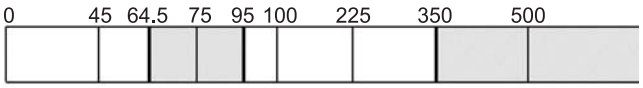

Fig. 2. Graphical representation of the rules

The graphical representation of the rules appears in Fig. 2, which is more understandable. The gray areas represent lethal values and white areas – non-lethal cases. The thin lines into the areas denote exceptions. To know more about exceptions you need to look at the rules.

The developed rule set was applied to other drugs of "General anesthetic compounds". The results are:

04 - > 2/10, 05 - > 0/1, 06 - > 5/5, 08 - > 3/7, 09 - > 8/13, 10 - > 10/12

where 04 - > 2/10 means: in the database there are total 10 instances on Drug 04 of which 2 are classified wrongly.

The best result is for Drug 04. It is a candidate that can be included to the developed rule set.

For Drug 05 there is 1 instance only, so can be ignored. For the remainder drugs seems likely to create special rules. As an example a rule was created for Drug 10 that has a simple structure:

if (Dose > 2000) then Lethal = yes else Lethal = no.

Applications

The applications of the rules are demonstrated on three examples.

Example 1

One wants to investigate effects of Drug 07 on an animal B (rabbit).To accomplish this, first estimate the safe borders for Dose. Examination of the rule set (Fig. 2) revealed that safe zones are expressed in the rules R1 and R3. Taking the safe value, say 60, apply it to the route IP. The rule R1 gives Lethal = yes. So the experiment is not safe. Trying another route, say PO, one can receive a desired result.

Example 2

Experiment is to be done on a dog (D). Find the possible higher value for safe border for Drug 01. The R4 gives the value 350. But on examining the exception to the rule one discovers that for the route PO the safe border can be increased till 500.

Example 3

Providing series of experiments on animals rabbit (B), cat (C) and dog (D) using Drug 02 with a safe dose interval (95 – 350). The rule R3 gives the following recommendations:

– for animals, rabbit and dog, the interval is safe for routes IP, IV, PO and SC, but for a cat only routes IV, PO and SC can be used.
– for a cat and a route IP doses greater than 150 are not recommended.

RESULTS AND DISCUSSION

The great amount of data on animal experiments contains a wide range of useful information that is important in education and research in pharmacology. In this paper results of some common statistical analysis of animal experimental data related with the total numbers of used compounds, animals, administration routes and animal actions, were considered. Based on these data it is also possible to provide more specific analysis, for example, to answer the question like "How many experiments were resulted with the Action on kidneys and which animal was used more frequently in them?" (The answer is: total 32 experiments, Dog is used in 14).

Another possibility for using knowledge drawing from experimental data is related with the problems of determining drug dosages. There are great numbers of possible animal actions observed in experiments (main categories appear in Appendix B). In some cases it is likely to take one action (single or categorical) and try to determine dose limits for the action values of type (yes/no). In our model we have used this approach – whether Toxicity action is lethal or non-lethal. In other situations binary values might not be desirable - multiple-valued attributes would to be used. For example, one might need to have for Toxicity more detailed values such as Lethal, Acute toxic, Chronic toxic, None. Then the model must be able to determine dose values for each of them.

The developed model is based on rules of easy understandability and applicability and can be used by researchers in planning new animal experiments and by students in educational processes. In this investigation the rules were designed for a restricted group of drugs, animals and routes because of absence of sufficient structural similarities in the dataset that is important for making common decisions. Creating different rule sets for specific component groups can be followed by designing specific software that covers them all. Such software will be very useful especially in education.

The total numbers of drugs, animals and administration routes that take place in animal experiments (and stored in the database) are 212, 7, and 5 respectively. The database contains generally a single record (rarely 2) for each used value of a triple "drug-animal-route". Because of minority of instances on meaningful components of data the application of classical Data Mining methods to whole dataset fails: no patterns, no visible shapes and borders. In such situations a searching process can go in restricted areas which usually have small dimensions. One approach to this is used in this investigation. First create a training dataset; including to it instances selected using specific methods. Methods include: seeing for informatively of units, estimating attribute values before an instance is selected, avoiding repetition in attribute values etc. Then see if there are some structural patterns in the training dataset (here a decision tree approach is used). If so, apply the found scheme to a whole dataset (which is also small) using it directly or using a set of rules constructed on its base. If all the instances (or at least a greatest part of them) are covered by the scheme, solution is found. If not, pruning (can be also called as generalization) stage starts. Creating exceptions to rules is one of such methods. For understandability and applicability reasons structures of exception are to be simpler not having, e.g., multiple exceptions to exceptions. In the present case the defined rule set is simple and easily applicable. When new experiments' results differing from existing data arise the rule set can be modified to cover them or a creation process can be started from the beginning.

REFERENCES

Babanli A, Gunes A and Ozturk Y. 2011. A computer experiment model to investigate the effects of drug dosage in animals, for use in pharmacological education and research. *Alternatives to Laboratory Animals* **39**: 519–25.

Barnes S D and Eltherington L G. 1973. *Drug Dosage in Laboratory Animals: A Handbook,* pp. 313, Berkley, CA, USA.

Borchard B E, Barnes C D and Elthington L G. 1990. *Drug Dosage in Laboratory Animals: A Handbook,* pp. 692, Telford Press, New Jersey, USA.

Boyer S. 2009. The use of computer models in pharmaceutical safety evaluation. *Alternatives to Laboratory Animals* **37**: 519–25.

Durieux P, Tringuart L, Colomber I, Nies J, Walton R T and Rajeswaran A. 2010. *The Cochrane Collaboration*, John Wiley & Sons.

Gunes A, Ozturk Y and Babanli A. 2009. An invivo database model for pharmocological and physiological dosage for experimental animals. *Computers in Biology and Medicine* **39** (7): 590–94.

Russell S and Norvig P. 2003. *Artificial Intelligence. A Modern Approach*. 2nd edn, pp. 649–73, New Jersey, USA, Pearson Education Inc.

Watton R, Dovey S, Harvey E and Freemantle N. 1999. Computer support for determining drug dose: Systematic review and meta-analysis. *British Medical Journal* **18**: 984.

Witten I H and Frank E. 2005. *Data Mining. Practical Machine Learning Tools and Techniques.* 2nd edn, pp.83–139. Elsevier, San Francisco, CA, USA, Morgan Kaufman Publications.