Machine Learning Analysis of the Relationship Between Changes in Immunological Parameters and Changes in Resistance to *Listeria monocytogenes*: A New Approach for Risk Assessment and Systems Immunology

Zhifa Liu,* Changhe Yuan,* and Stephen B. Pruettt†¹

*Department of Computer Science and Engineering, Bagley College of Engineering, and †Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi 39762

¹To whom correspondence should be addressed at Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, P.O. Box 6100, MS 39762. Fax: (662) 325-8884. E-mail: Pruettt@cvm.msstate.edu.

Received January 11, 2012; accepted May 22, 2012

No method has been reported to predict, even approximately, the impact of mild-to-moderate changes in several immunological parameters on resistance to infection. The ability to make such predictions would be useful in risk assessment. In addition, equations that predict host resistance on the basis of changes in components of a complex biological system (the immune system) would fulfill one of the major goals of systems biology. In this study, multiple machine learning classification methods were used to predict the effects of a series of drugs and chemicals on host resistance to *Listeria monocytogenes* in mice on the basis of changes in several holistic immunological parameters. A data set produced under the sponsorship of the National Toxicology Program (NTP) was used in this study. The NTP data set was found to have a high percentage of missing data and to be noisy (probably due to the intrinsically stochastic nature of immune responses). Data preprocessing steps were used to mitigate these problems. In evaluating the machine learning classifiers, we first randomly partitioned the NTP data set into 10 subsets. Each time, we used nine subsets of the data to train the machine learning classifiers, and the remaining single subset to predict outcomes with regard to host resistance. This process was repeated until all 10 combinations of the 9-1 split of the subsets have been tested. The best of the classifiers predicted host resistance outcome correctly for 94.7% of cases, a result which indicates it is possible to identify mathematical expressions that will be useful for risk assessment and to establish a basis for systems immunology.

Key Words: missing data imputation; discretization; classification; cross validation; resistance to infection; immunotoxicity; immunosuppression; systems immunology; immunological parameters.

Quantitative relationships between immune defense mechanisms and host resistance to infection or cancer have not been well characterized. It is clear that substantial suppression (Chan and Allen, 2002; Fung *et al.*, 2001; Tufano, 2002; Zandman-Goddard and Shoenfeld, 2003) or elimination

(Nathwani *et al.*, 2005) of a major immunological function or cell type decreases resistance to particular infections or cancer. However, the effects of lesser suppression of several immunological functions on host resistance cannot presently be accurately predicted. The most comprehensive effort till date in this regard indicated that mild or moderate changes in single immunological parameters, even those known to be important in resistance to a particular pathogen, were not good predictors of host resistance to that pathogen (Luster *et al.*, 1993). This probably reflects the complexity and redundancy of some immune functions as well as the complexity of drug, chemical, and environmental effects on them.

Fully understanding the immune system will require understanding quantitative as well as qualitative relationships between immunological parameters and host resistance. This has become evident in a number of situations that are relevant to human health. For example, there is abundant and consistent evidence that neuroendocrine stress responses can suppress a number of immunological parameters (Herbert and Cohen, 1993). Drugs of abuse as well as occupational or environmental exposure to chemicals are also associated with suppression of immunological parameters (Friedman and Lawrence, 2002). In some cases, suppression of host resistance has also been documented (Cohen *et al.*, 1998; Vedhara *et al.*, 1999). However, the degree of increased risk of infection or cancer caused by a particular amount of suppression of particular immunological parameters is not known. Developing an approach to accomplish this would have practical implications for regulatory agencies (e.g., FDA or EPA), which must make risk-benefit decisions on the basis of limited data that often includes some immune function measures but seldom includes thorough assessments of host resistance or incidence of infection.

In a previous publication, we presented results demonstrating that only about 60% of the decrease in resistance to B16F10 melanoma in mice treated with various doses of an

immunosuppressant could be accounted for by multivariate regression modeling using parameters that encompass the major resistance mechanisms (nitric oxide production by macrophages, antigen-specific cytotoxicity by T cells, and activity of natural killer [NK] cells) (Keil *et al.*, 2001). Only accounting for 60% of the change in host resistance caused by a known immunosuppressive drug suggests either that other (unknown) mechanisms are also important in host resistance, that the modeling methodology was inadequate, or both. With regard to modeling, logistic regression, which was used in our previous study, has been widely used in ecology and other fields to describe and predict the role of many explanatory variables in changes observed in a single dependent variable. However, this method has limitations. It only takes into account linear relationships between explanatory variables and the categorical dependent variable, and it does not account for more complex relationships in which changes in one parameter may be compensated by lack of change or by change in the opposite direction of another parameter.

An artificial intelligence-based pattern recognition system has the potential to overcome some of the limitations of multivariate modeling. In particular, machine learning classifiers can detect patterns of interactions or quantitative relationships present in a data set that would not be detectable by comparing graphical representations of the results. Furthermore, some of these machine learning methods do not assume any particular relationship (linear, nonlinear, etc.) between particular immunological parameters and host resistance. The purpose of this study was to apply machine learning methods to accurately predict resistance to *Listeria monocytogenes* infection in mice based on changes in several immunological parameters caused by exposure to drugs and chemicals investigated by the National Toxicology Program (NTP). Due to the noise and missing data present in the data set, it was necessary to perform several statistical data preprocessing steps before analysis with several machine learning classifiers. The utility of these steps in this study suggests that they will be useful in the quantitative analysis of many data sets in which there are similar problems, which would probably include most immunological data sets, at least to some degree. The analysis of the preprocessed data showed that the machine learning methods skillfully predicted the outcome with regard to mortality after infection with *L. monocytogenes*.

These results demonstrate for the first time that it is feasible to quantitatively estimate the effects of changes in one or more immunological parameters on resistance to bacterial infection in mice. This has implications with regard to risk assessment and fills a gap in our current understanding of the immune system by quantifying the relationships between changes in immunological parameters and changes in host resistance to infection. In addition, the set of data preprocessing methods described in this study should be useful for analysis of a wide variety of biological data sets, which often have problems with noise and missing values. These studies are also relevant to the emerging field of systems immunology because they provide mathematical algorithms to explain and predict a complex function (host

resistance to infection) on the basis of changes in individual components responsible for this function. Developing such algorithms is a key goal in systems biology, and this is further delineated in the Discussion section.

MATERIALS AND METHODS

Data for these studies. This study takes advantage of a unique data set acquired under the auspices of the NTP. This data set is well suited for the purpose of this study because results were obtained using Good Laboratory Practice (GLP) standards, and many of the immunological parameters evaluated have been validated in interlaboratory studies (Luster *et al.*, 1988). The origins of the NTP data set can be traced back to data obtained and placed in a database by Gayla Pait for two foundational studies (Luster *et al.*, 1992; Luster *et al.*, 1993) on the relationship between immune function parameters and host resistance. Technical details for the protocols used to evaluate immunological parameters are provided in publications cited in this section and on the NTP Web site (<http://ntp.niehs.nih.gov/?objectid=12EDA31C-AE8F-F3CD-3C85075C0351EC2E>). The mice used in most studies were of the same age (8–10 weeks), gender (female), and strain (C57Bl/6 × C3H/HeN F1), but different strains were used in a few studies. The purpose of NTP immunotoxicity studies was to identify compounds to which people are exposed environmentally, occupationally, or as therapeutic agents that have the potential to adversely affect the immune system.

It should be noted that the immunological parameters evaluated in this data set are generally the most holistic parameters relating to a particular immune cell type or function. Thus, NK cell cytolytic function, not molecular endpoints related to this function, is included in the data set. The inclusion of such holistic parameters was intentional and was based on the idea that these parameters require several molecular events and/or cell types and thus are more likely to permit detection of processes that are disrupted by immunotoxic substances. The detailed data set used in this study was kindly provided by Dr Dori Germolec (NTP, National Institute of Environmental Health Sciences). A sample of the actual NTP data is shown in Table 1 for one chemical (benzidine [BZ]) that was tested. There are totally 42 drugs and/or chemicals. This does not constitute the whole NTP data set, but it represents the data available in spreadsheet format that was necessary for the type of analysis carried out in this study. In future studies, we plan to use the machine learning methods that performed best in this study to evaluate the whole NTP data set. Some of the 42 drugs and chemicals in the data set were eliminated from analysis due to excessive missing values or outliers. For several chemicals, the results of the NTP studies have been published (Guo *et al.*, 2000; Holsapple *et al.*, 1988; Karrow *et al.*, 2000; Karrow *et al.*, 2003; LeVier *et al.*, 1994; Sikorski *et al.*, 1989). For most chemicals, three different doses were used and there was a vehicle control group. Exceptions to this were silicone, nickel sulfate, ginseng, formaldehyde, and toluene. Silicone was tested at four different doses. Nickel sulfate was tested at two doses, and ginseng, formaldehyde, and toluene were tested at only one dose. Therefore, this subset of the NTP data set has 158 data cases (a case equals one dose of one chemical as well as vehicle controls in all experiments).

The immunological parameters captured in the NTP data set were body weight (BW), spleen weight (SW), spleen cellularity (SCell), SW:BW ratio (S:B), percentage of T cells, CD4⁺, CD8⁺, and Ig⁺ cells in the spleen (Thy 1.2⁺, CD4⁺, CD8⁺, Ig⁺), thymus weight (TW), thymus cellularity, plaque-forming cells per spleen (PFC/spleen) and per 10⁶ spleen cells (PFC/10⁶) in response to sheep red blood cells, response of spleen cells to concanavalin A (ConA), phytohemagglutinin (PHA), lipopolysaccharide (LPS), nonhistocompatible mouse spleen cells treated chemically to prevent cell division (mixed lymphocyte reaction [MLR]), delayed hypersensitivity response to a known sensitizer, cytotoxic activity of NK cells against YAC-1 target cells (NK), lysis of nonhistocompatible target cells by cytotoxic T lymphocytes (CTL), and white blood cell (WBC) concentration in blood. Resistance to infection was evaluated using *L. monocytogenes* administered intraperitoneally. The dose of *Listeria* used was less than the LD₅₀ and typically less than 20% of mice in the control group died.

TABLE 1
Data from the NTP Data Set for One Chemical, BZ

Variable no.	Variable	Case 1	Case 2	Case 3	Case 4
		Dose 1, BZ	Dose 2, BZ	Dose 3, BZ	Dose 4, BZ
1	Dose (mg/kg)	0.00	11.00	22.00	44.00
2	Body weight (g)				
3	SW (mg)				
4	Spleen:body				
5	TW (mg)				
6	Thymus:body				
7	Cells/spleen ($\times 10^7$)	24.10	26.60	29.70	21.20
8	WBC ($\times 10^6$ /ml)				
9	Thy1.2 (% spleen)	40.00	35.00	43.00	38.00
10	CD4				
11	CD8				
12	Ig+ (% spleen)				
13	PFC/ 10^6	2332.00	1953.00	2101.00	1751.00
14	PFC/spleen	61.00	48.00	52.00	49.00
15	LPS (spleen)				
16	NK cell activity	16.00	17.00	19.00	9.00
17	MLR (spleen)	106.00	81.00	68.00	54.00
18	DHR (spleen)	3.60	3.60	3.00	2.40
19	CTL (spleen)				
20	PHA (spleen)	70.00	69.00	62.00	37.00
21	ConA (spleen)	164.00	134.00	95.00	70.00
Target	Mortality after <i>Listeria</i> challenge	0.20	0.60	0.60	1.00

Note. These are raw values, and data are missing for one or more immunological parameters, which is typical in this data set. Suppression of some of these parameters as the dose of BZ increases is associated with an increase in mortality from 0.2 (20%) to 1.0 (100%).

Feature selection. Several of the aforementioned immunological parameters were removed before the machine learning analysis because they either are not considered as good indicators of resistance to infection or are strongly correlated with another attribute. For example, BW ratio is not a good indicator of resistance because it did not vary much with dose. We also excluded SW, SW to BW ratio, TW, and TW to BW ratio for the same reason. For the remaining attributes, we conducted a pairwise collinearity study (Figure 1). The regression analysis of CD4⁺ and CD8⁺ shown in Figure 1 indicates that these two attributes are strongly linearly correlated. This correlation is consistent with the well-known homeostatic control of the CD4⁺ to CD8⁺ T-cell ratio. This regression analysis was only done for pairs without missing data. We also did the collinearity analysis for all the attribute pairs after missing values were imputed (see Missing Data Imputation section). The correlation graphs and coefficient matrix are shown in Figure 2 and Table 2, respectively. The new regression plot between CD4⁺ and CD8⁺ shows similar results; the correlation coefficient between these two attributes is as high as 0.983. Therefore, to avoid collinearity, we included only one of these two immunological parameters in the analysis, CD4⁺.

Outlier detection. Even though the NTP data set was obtained using GLP standards and most of the immunological procedures have been validated by inter-laboratory studies (Luster *et al.*, 1988), the data set still contains a significant amount of noise. This is not unusual for many immunological parameters and probably reflects the substantial role of stochastic factors in immunological responses. Nevertheless, some values in the data set are clearly out of the range of biologically reasonable values and probably reflect technical mistakes or assay errors rather than normal biological variation. For example, most of the values of the parameter PFC/ 10^6 are less than 3000 in the NTP data set, but the maximum value of the attribute is close to 6000. Such extreme values are considered as outliers for the purpose of this machine learning analysis because they would adversely affect the capability of machine learning classifiers in detecting useful association patterns or mechanisms from the data.

Outlier detection techniques can be used to identify and remove outlier values to improve the quality of the data. One commonly used method is called

Boxplot (Tukey, 1977), a graphic display method for indicating outliers. Boxplot marks outliers based on the quartile measures. Let Q_1 and Q_3 be the first and third quartiles, respectively. Boxplot defines an outlier to be any value outside of the following range:

$$[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$$

Figure 3 shows the boxplots of several parameters, including PFC/ 10^6 , PFC/spleen, ConA, and SCell, which provide a clear display of the outlier values. However, we did not rely solely on boxplot in determining the outliers. For any outlier detected, we also used expert knowledge to make the final ruling on whether it is a true outlier or not. One of us (S.B.P) has considerable experience measuring these parameters in mice of the same age, gender, and strain as the ones used in the NTP studies (Glover and Pruett, 2009; Han *et al.*, 1993; Han and Pruett, 1995; Hebert and Pruett, 2003; Pruett *et al.*, 2004) and, therefore, is in a position to make expert judgments regarding the range of values typical for these parameters. In most cases, outliers identified by the boxplot method were confirmed as outliers, but in a few cases, the observed values were within a range that had been seen relatively frequently in other data sets and these were not classified as outliers. Any case (one chemical at a particular dose) that contained true outlier values was removed from the final data set. About 15% of the cases were deleted, yielding a remaining 133 cases after outlier detection and deletion.

Missing data imputation. Unfortunately, most of the machine learning methods used in this study are not tolerant of missing data. One option to address missing data is to delete cases that include parameters with missing data, leaving only cases without missing data. However, in the NTP data set, this would leave essentially no cases because all cases contain at least one parameter with missing value (Supplementary table S1). The percentage of missing values ranges from 5% to as high as 75%. Figure 4 shows the percentage of missing values for all the attributes (defined here as immunological parameters being evaluated for predictive value). Even pattern recognition methods that are capable of handling missing values may

FIG. 1. Regression analysis between $CD4^+$ and $CD8^+$ cells for known values only (not imputed values). The raw values for $CD4^+$ and $CD8^+$ (percent of cells in spleen) are shown. The results are consistent with the well-known stability of the $CD4^+$ to $CD8^+$ ratio. The collinearity between these parameters would have an adverse effect on the machine learning methods, so $CD8^+$ was eliminated from the data set.

be also adversely and unpredictably affected by the presence of this many missing values.

Various methods have been developed for dealing with missing values. Case deletion, also known as complete case analysis, is a simple technique, which removes all data cases and attributes that contain missing values. This method is not applicable to the NTP data set because results for all the chemicals tested

contain missing values. Data imputation methods are commonly used to handle missing values by replacing them with plausible values based on some kind of statistical estimation technique. The mean imputation and k -nearest neighbor (KNN) methods are two examples. Mean imputation replaces a missing value of an attribute (immunological parameter) with the mean of all known values of the same attribute. That is, a missing value of an attribute A_k will be replaced by

$$\bar{x}_k = \sum_{i: x_{ik} \neq \text{nil}} \frac{x_{ik}}{n_k},$$

where x_{ik} 's are the known values of this attribute and n_k is the number of known values of A_k .

KNN (Aha and Kibler, 1991) replaces a missing value in a data point with the most frequent value or the mean value of the same attribute in the k -nearest data cases (drug or chemical used and dose). The distance between two data cases is calculated according to a distance measure. KNN first divides the data cases into two sets, D_m and D_c , with D_m being the set containing missing values and D_c being the set without missing values. For each missing value of a data case in D_m , KNN first finds the k -nearest data cases from the complete data set D_c . If the missing value is categorical, it is replaced by the majority value of the same attribute in those k data cases. If continuous, the missing value is replaced by the mean value of the k data cases.

KNN is not applicable to the NTP data set either because there are missing values for all the chemicals. As a result, the D_c set would contain no data cases. We present an improved version of the KNN method called Sparse-KNN to handle such data sets. Sparse-KNN no longer divides the data set into two subsets. Instead, it directly finds the k -nearest data cases using a distance measure that is robust in face of missing values. Because the NTP data set only has continuous attributes, we define the distance between the i th and j th data cases as

$$D_{ij} = \frac{\sqrt[q]{\sum_{k: x_{ik} \neq \text{nil}, x_{jk} \neq \text{nil}} |x_{ik} - x_{jk}|^q}}{K_{ij}}$$

where K_{ij} is the number of attributes that have both x_{ik} and x_{jk} known. The parameter q can be set to different values to represent different distance measures. For example, if $q = 1$, the measure represents the Manhattan distance. If

FIG. 2. Collinearity analysis between the attributes after missing values was imputed. Normalized values for each parameter for all drug and chemical doses were plotted to reveal collinear relationships between parameters, which would adversely affect machine learning if not removed. For example, the second plot in the first column represents the relationship between SCell and WBC. The only strongly collinear parameters were $CD4^+$ and $CD8^+$.

TABLE 2
Correlation Coefficient Matrix Between the Attributes after Missing Values Are Imputed

	SCell	WBC	Thy1.2	Ig+	PFC	PFC/SP	LPS	NK cell	MLR	PHA	ConA	DHR	CD4	CD8	CTL	Listeria
SCell	1	0.287 8E-04	0.794 <.0001	0.763 <.0001	0.372 <.0001	0.689 <.0001	0.375 <.0001	0.003 0.974	0.502 <.0001	0.556 <.0001	0.469 <.0001	0.226 0.009	0.786 <.0001	0.784 <.0001	0.82 <.0001	-0.21 0.018
WBC	0.287 8E-04	1	0.433 <.0001	0.233 0.007	0.191 0.028	0.156 0.074	0.221 0.011	0.378 <.0001	0.216 0.013	0.183 0.036	0.189 0.03	0.339 <.0001	0.289 8E-04	0.297 5E-04	0.289 8E-04	0.01 0.907
Thy1.2	0.794 <.0001	0.433 <.0001	1	0.747 <.0001	0.336 <.0001	0.494 <.0001	0.461 <.0001	0.07 0.425	0.407 <.0001	0.523 <.0001	0.409 <.0001	0.214 0.014	0.73 <.0001	0.758 <.0001	0.555 <.0001	-0.2 0.022
Ig+	0.763 <.0001	0.233 0.007	0.747 <.0001	1	0.405 <.0001	0.637 <.0001	0.358 <.0001	0.162 0.064	0.395 <.0001	0.329 1E-04	0.459 <.0001	0.147 0.093	0.853 <.0001	0.823 <.0001	0.609 <.0001	-0.22 0.013
PFC	0.372 <.0001	0.191 0.028	0.336 <.0001	0.405 <.0001	1	0.709 <.0001	0.02 0.818	-0.01 0.88	0.271 0.002	0.145 0.097	0.098 0.266	0.304 4E-04	0.344 <.0001	0.282 0.001	0.42 <.0001	0.033 0.707
PFC/SP	0.689 <.0001	0.156 0.074	0.494 <.0001	0.637 <.0001	0.709 <.0001	1	0.119 0.173	-0.11 0.19	0.347 <.0001	0.374 <.0001	0.36 <.0001	0.288 8E-04	0.703 <.0001	0.635 <.0001	0.746 <.0001	-0.15 0.082
LPS	0.375 <.0001	0.221 0.011	0.461 <.0001	0.358 <.0001	0.02 0.818	0.119 0.173	1	0.109 0.212	0.514 <.0001	0.389 <.0001	0.505 <.0001	-0.03 0.704	0.44 <.0001	0.494 <.0001	0.303 4E-04	0.024 0.782
NK cell	0.003 0.974	0.378 <.0001	0.07 0.425	0.162 0.064	-0.01 0.88	-0.11 0.19	0.109	1	0.125 0.154	0.022 0.806	0.086 0.327	0.389 <.0001	-0.02 0.824	-0.01 0.875	-0.01 0.879	0.081 0.357
MLR	0.502 <.0001	0.216 0.013	0.407 <.0001	0.395 <.0001	0.271 0.002	0.347 <.0001	0.514 <.0001	0.125 0.154	1	0.465 <.0001	0.526 <.0001	0.144 0.1	0.523 <.0001	0.566 <.0001	0.541 <.0001	-0.07 0.439
PHA	0.556 <.0001	0.183 0.036	0.523 <.0001	0.329 1E-04	0.145 0.097	0.374 <.0001	0.389 <.0001	0.022 0.806	0.465 <.0001	1	0.336 <.0001	0.286 9E-04	0.48 <.0001	0.506 <.0001	0.539 <.0001	-0.2 0.023
ConA	0.469 <.0001	0.189 0.03	0.409 <.0001	0.459 <.0001	0.098 0.266	0.36 <.0001	0.505 <.0001	0.086 0.327	0.526 <.0001	0.336 <.0001	1	0.087 0.322	0.613 <.0001	0.621 <.0001	0.527 <.0001	-0.21 0.014
DHR	0.226 0.009	0.339 <.0001	0.214 0.014	0.147 0.093	0.304 4E-04	0.288 8E-04	-0.03 0.704	0.389 <.0001	0.144 0.1	0.286 9E-04	0.087 0.322	1	0.145 0.097	0.131 0.134	0.304 4E-04	0.103 0.239
CD4	0.786 <.0001	0.289 8E-04	0.73 <.0001	0.853 <.0001	0.344 <.0001	0.703 <.0001	0.44 <.0001	-0.02 0.824	0.523 <.0001	0.48 <.0001	0.613 <.0001	0.145 0.097	1	0.983 <.0001	0.77 <.0001	-0.32 2E-04
CD8	0.784 <.0001	0.297 5E-04	0.758 <.0001	0.823 <.0001	0.282 0.001	0.635 <.0001	0.494 <.0001	-0.01 0.875	0.566 <.0001	0.506 <.0001	0.621 <.0001	0.131 0.134	0.983 <.0001	1	0.737 <.0001	-0.32 1E-04
CTL	0.82 <.0001	0.289 8E-04	0.555 <.0001	0.609 <.0001	0.42 <.0001	0.746 <.0001	0.303 4E-04	-0.01 0.879	0.541 <.0001	0.539 <.0001	0.527 <.0001	0.304 4E-04	0.77 <.0001	0.737 <.0001	1	-0.24 0.006
Listeria	-0.21 0.018	0.01 0.907	-0.2 0.022	-0.22 0.013	0.033 0.707	-0.15 0.082	0.024 0.782	0.081 0.357	-0.07 0.439	-0.2 0.023	-0.21 0.014	0.103 0.239	-0.32 2E-04	-0.32 1E-04	-0.24 0.006	1

Note. This table provides the numerical results of the collinearity analysis shown in Figure 2.

$q = 2$, it is the Euclidean distance. We used the Euclidean distance in this study. Essentially, this measure calculates the distance between two data cases by only considering attributes with both values known. The Sparse-KNN then works as follows. To fill in a missing value for a data case, it first finds the k -nearest data cases based on the modified distance measure. It then replaces the missing value with the mean value of the same immunological parameter in these k neighbors. The method can also easily be generalized to categorical attributes.

We applied both the mean value imputation and Sparse-KNN methods to the NTP data set in order to compare their performance. We set k to be 5 in this study. Before applying the value imputation methods, we normalized the values for all of the immunological parameters to be within the range of 0 and 1 before applying the value imputation methods. The normalization is necessary because the original values of the parameters typically have different scales and may affect the distance measure disproportionately.

Discretization of the class variable. We formulated the prediction of resistance to infection in the NTP data set as a classification problem. The resistance to infection is indicated by the mortality rate of the mice in each experiment. We discretized the mortality rate into several categories in three different ways based on expert knowledge. In the first way, we divided the mortality rates into two groups based on a threshold mortality rate, with the first group less than or equal to the threshold and the second group greater than the threshold. The first group was considered as the low-mortality population, whereas the second group is the high-mortality population. Two different threshold mortality rates were tested in this study: 30% (Type_2_0.3) and 50% (Type_2_0.5). The second way of discretization divides mortality rates into three groups by using two dividing points: 30% and 70%.

The three intervals correspond to low-(0–30%), medium-(30–70%), and high-mortality (70–100%) populations, respectively (Type_3). Finally, we also considered a four-category discretization in which 25, 50, and 75% are the dividing points (Type_4). The four intervals correspond to low- (0–25%), medium low- (25–50%), medium high- (50–75%), and high-mortality (75–100%) populations, respectively. Table 3 provides a summary of the different discretization criteria of the mortality rate. This allows prediction of survival of mice to the level of two, three, or four categories. These categorized (discretized) values allow the formulation of predicting the resistance level as a classification problem.

Discretization of the attributes (immunological parameters). Many machine learning classification methods only deal with categorical attributes. Therefore, we also discretized the attributes of the NTP data set into categorical variables. Discretizing all the attributes manually would require an unacceptable amount of time. Therefore, we used automatic discretization methods. There are two types of such methods: unsupervised and supervised discretization methods. Unsupervised methods discretize the attributes without considering the class variable. Two commonly used methods are equal-width interval discretization and equal-frequency interval discretization. Both of these sort the observed values of a continuous attribute and naively divide the attribute into k equal-width or equal-frequency bins, where the number of bins is a parameter supplied by the user.

Supervised methods discretize the attributes so as to maximize the effectiveness of these attributes in predicting the class variable. There are two popular methods in this category: minimum description length (MDL)-based discretization method (Fayyad and Irani, 1993) and ChiMerge (Kerber, 1992). The MDL-based discretization method selects bin boundaries for discretization that

FIG. 3. Boxplot of attribute PFC/10⁶. The dots indicate outlier values. The central horizontal line represents the mean and the lower and upper border of the box represent the first and third quartile values. The error bars were determined using the formula described in Materials and Methods section, and data cases outside these bars are shown by open circles and are considered outliers.

minimize the class information entropy of candidate partitions. Given a set of data cases S , an attribute A , and a partition boundary T (S_1 and S_2 are the resulting two subsets), the class information entropy of the partition is defined as (Fayyad and Irani, 1993)

$$E(A, T; S) = \frac{|S_1|}{|S|} \text{Ent}(s_1) + \frac{|S_2|}{|S|} \text{Ent}(s_2)$$

The method uses the MDL principle to determine a stopping criterion for their recursive discretization strategy. Let N be the number of instances in the set S , and k_i be the number of class levels in each subset S_i . The partition induced by a cut point T within a set of values S will be accepted if and only if

$$\text{Gain}(A, T; S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A, T; S)}{N},$$

where

$$\text{Gain}(A, T; S) = \text{Ent}(S) - E(A, T; S) \text{ and}$$

$$\Delta(A, T; S) = \log_2(3^k - 2) - [k * \text{Ent}(S) - k_1 * \text{Ent}(S_1) - k_2 * \text{Ent}(S_2)]$$

ChiMerge is a method that uses the χ^2 test to determine whether the relative class frequencies of two adjacent intervals are distinctly different or whether they are similar enough so that merging them into a single interval is justified. The algorithm includes two stages: an initialization stage and a bottom-up merging stage. It starts by sorting the observed values of a continuous attribute and building the initial discretization with each example put into its own interval. In the merging stage, the intervals are recursively merged together until all pairs of intervals have χ^2 values exceeding a threshold. The formula for computing the χ^2 value is as follows:

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

where k is the number of class, m is the number of intervals being compared, A_{ij} is the number of examples in the i th interval and j th class; and E_{ij} is the expected frequency of A_{ij} , defined as $\sum_{j=1}^k A_{ij} * \sum_{i=1}^m A_{ij} / \sum_{j=1}^k \sum_{i=1}^m A_{ij}$.

The threshold for the χ^2 value is typically set based on a desired significance level. The smaller the threshold value, the smaller the number of intervals and the larger the widths for the intervals.

We applied both unsupervised and supervised discretization methods to the NTP data set. For the unsupervised method, we used the equal-width interval discretization method. For the supervised method, we tried both the MDL-based method and ChiMerge. The results show that the MDL-based method discretized most of the host resistance data into only one interval for this data set; it discretized several other attributes into similar bins as ChiMerge though. Figure 5 shows a comparison of these two discretization methods for the attribute Thy1. We, therefore, selected the ChiMerge method to discretize the data for this study. However, we noticed that, if we set the same threshold for the different attributes in ChiMerge, some attributes would have very few intervals and others would have dozens of intervals. Because we only have 133 data cases in the processed NTP data set, such a large number of intervals is unjustified. We, therefore, adjusted the threshold value manually so that each attribute has 2–7 intervals. Table 4 provides the detailed discretization criteria for several sample attributes.

Cross-validation accuracy tests with several machine learning classifiers.

After all the data preprocessing steps, we finally had a data set that was ready to be analyzed using standard machine learning classifiers. We applied a variety of classification methods implemented in the WEKA tool (Hall *et al.*, 2009) to the NTP data set, including naive Bayes (McCallum and Nigam, 1998), KNN (Aha and Kibler, 1991), C4.5 (Quinlan, 1993), bagging (Breiman, 1996), boosting (Friedman *et al.*, 2000), support vector machine (SVM) (Vapnik, 1995), and multilayer perceptron (neural network) (Pal and Miltra, 1992). Bagging and boosting are ensemble classification methods built on top of other classifiers. In this study, we used a decision tree method as the base classifier for bagging and boosting. We selected these machine learning methods for this study because the methods are considered to be representative classifiers in the machine learning field. A tenfold cross validation was used in our experiments. The basic idea is to randomly divide the complete data set into 10 sets. Each time, we used nine sets of the data to train the machine learning classifiers and the remaining set to evaluate the accuracy of the classifiers. This process was repeated for each set of the data. The average accuracies of the results over the 10 experiments are shown in Table 5.

Because several attributes have as many as 70–80% of missing values, it is a legitimate concern that filling in those missing values using the Sparse-KNN method may introduce bias. Therefore, we also conducted an experiment by removing all the features with high missing rates. The changes in accuracy results are shown in Table 6. In addition, we removed the values deemed to be outliers by both mathematical methods and domain experts. We conducted another experiment for data sets with those outlier values retained. The changes in accuracy results are shown in Table 7. The missing value imputation and attribute discretization step follow the same rules mentioned in Table 5.

RESULTS

The data set on which these analyses were based can be found as supplemental data associated with this manuscript online. This data set includes only the portion of the NTP data that was available in spreadsheet format a few years ago when the data were obtained. The data show the effects of 42 drugs or chemicals on 21 immunological parameters and on resistance to *L. monocytogenes* infection (as indicated by percent mortality). In most cases, the groups included a vehicle control group and three doses of the drug or chemical. The parameters represented in the NTP database include parameters important in humoral immunity (PFC and Ig+), cell-mediated immunity (MLR, DHR,

FIG. 4. Fraction of cases with missing values for all the immunological parameters. The total number of cases (a case is one dose of a chemical) divided by the number of cases for which data are available for particular immunological parameters are plotted. Thus, data for CD8⁺ were missing in more than 70% of cases, but data for PFC/10⁶ were missing for only 10% of cases.

TABLE 3
Criteria for Discretizing the Class Variable

Type	Type_2_0.3	Type_2_0.5	Type_3	Type_4
Dividing points	0.3	0.5	0.3, 0.7	0.25, 0.5, 0.75

Note. Type_2_0.3: Discretize the class variable into two intervals: [0–0.3] and [0.3–1];

Type_2_0.5: Discretize the class variable into two intervals: [0–0.5] and [0.5–1];

Type_3: Discretize the class variable into three intervals: [0–0.3], [0.3–0.7], and [0.7–1];

Type_4: Discretize the class variable into four intervals: [0–0.25], [0.25–0.5], [0.5–0.75], and [0.75–1].

and CTL), and innate immunity (NK and WBC). A few of the parameters are no longer used much in immunology research (e.g., response to ConA and PHA mitogens), and their value is somewhat limited, but they nevertheless would reveal certain types of effects that would inhibit immune responses. Table 1 illustrates results for one chemical, BZ, and several features of these results are representative for other chemicals and drugs included in this data set. First, it is evident that several of the immunological parameters evaluated for some drugs and chemicals were not assessed for BZ, leading to several missing values. However, the data that were obtained are informative

FIG. 5. A comparison of the MDL-based and ChiMerge discretization methods on the attribute Thy1.2. The method of discretization can influence the predictive capability of machine learning algorithms. Two supervised methods of discretization were compared for several immunological parameters, and results for one of them (Thy 1.2 positive cells, which are T cells). This involves placing a particular immunological parameter for a particular drug or chemical dose or control into one of two bins depending on its numerical value. For example, a high value for an immunological parameter would be placed in bin 2 and a low value in bin 1 are shown here. Both methods discretize the attribute into two categories (1 and 2). The overlapped parts of the two lines show that MDL and ChiMerge agree on the discretized values for most of the data cases (122 out of 133 observations). This illustrates that two automated discretization methods give similar results, suggesting that automated discretization is suitable for use in this machine learning analysis.

TABLE 4
Criteria for Discretizing Several Example Attributes Used by ChiMerge

Attribute	Value 1	Value 2	Value 3	Value 4	Value 5
SCell	[0–17.9] ^c	(17.9–∞)			
WBC	[0–7.4]	(7.4–8.2]	(8.2–∞)		
Thy1.2	[0–53.8]	(53.8–∞)			
Ig+	[0–62.6]	(62.6–88.3]	(88.3–∞)		
PFC/10 ⁶	[0–1434.1]	(1434.1–1456.84]	(1456.8–∞)		
PFC/spleen	[0–239.7]	(239.7–268.3]	(268.3–276.1]	(276.1–∞)	
LPS	[0–15.7]	(15.7–18.0]	(18.0–43.0]	(43.0–134.6]	(134.6–∞)

^cThese discretization intervals were automatically generated by ChiMerge based on user-specified threshold values. “[“and “]” mean that the end values are included in an interval, while “(“and “)” mean not included.

and suggest that moderate, dose-responsive suppression of several immunological parameters such as NK cell activity, MLR, DHR, PHA, and ConA is associated with a substantial increase in mortality from 20% in control animals treated with vehicle to 100% in animals treated with the highest dose of BZ. Figure 6 illustrates the effects of some of the other drugs/chemicals on selected immunological parameters thought to be important with regard to mortality following challenge with *L. monocytogenes*. As noted previously by others (Luster *et al.*, 1993), there were no drugs or chemicals that decreased resistance to *Listeria* without also suppressing one or more immunological parameter. This suggests that the immunological parameters selected for the NTP studies include a representation of most of the major immunological functions that contribute to host resistance. In the examples shown in Figure 6, dose-responsive suppression is noted for some immunological parameters, and this is generally associated with increased mortality from *Listeria* infection. However, in two cases shown in Figure 6,

gallium arsenide and dideoxyadenosine, the compound being tested actually decreased mortality at high doses. Both of these compounds are directly toxic to bacteria as well as mammalian cells (Beskid *et al.*, 1981; Burns *et al.*, 1991), so increased survival is possibly due to the bactericidal effects of the compounds. Among the 45 drugs and chemicals in this data set, three drugs caused this type of response, the two already named plus morphine. Morphine induces a potent and long-lasting elevation of endogenous glucocorticoids, which is normally considered to be immunosuppressive, but it substantially increases the number of neutrophils, which may enhance host resistance (Freier and Fuchs, 1994; Pruett *et al.*, 2008). For the other chemicals and drugs, survival was decreased in a general dose-dependent manner. This is difficult to see on Figure 6 for dimethylbenzanthracene because the mortality for the control and lowest dose was 0, so only two bars appear, and they represent the mid and high doses. In this case, mid and high dose effects were not strictly dose responsive although there was a

TABLE 5

Cross-Validation Accuracy for Predicting % Mortality for the Classifiers in Data Sets Produced by Four Combinations of Imputation and Discretization Methods

Classifier	Type_2_0.3	Type_2_0.5	Type_3	Type_4
Missing value method: Mean imputation				
Discretization method: Bin-5				
Naive Bayes	0.568182	0.568182	0.530303	0.439343
KNN	0.674242	0.704545	0.628788	0.44697
C4.5	0.666667	0.795455	0.515152	0.454545
Boosting	0.651515	0.765152	0.58333	0.416667
Bagging	0.613636	0.803030	0.545455	0.515152
SVM	0.712121	0.704545	0.621212	0.431818
Multilayer perceptron	0.659091	0.696970	0.590909	0.462121
Discretization method: ChiMerge				
Naive Bayes	0.833330	0.924242	0.742424	0.666667
KNN	0.772727	0.795455	0.659091	0.568182
C4.5	0.772727	0.893939	0.628788	0.590909
Boosting	0.810606	0.886364	0.742424	0.689394
Bagging	0.825758	0.878788	0.712121	0.606061
SVM	0.613636	0.696970	0.621212	0.439394
Multilayer perceptron	0.787879	0.856061	0.666667	0.643939
Missing value method: SKNN-5				
Discretization method: Bin-5				
Naive Bayes	0.704545	0.704545	0.636364	0.522727
KNN-5	0.719697	0.757576	0.666667	0.537879
C4.5	0.704545	0.696970	0.689394	0.469697
Boosting	0.734848	0.742424	0.681818	0.515152
Bagging	0.742424	0.719697	0.643939	0.500000
SVM	0.689394	0.742424	0.575758	0.401515
Multilayer perceptron	0.734848	0.734848	0.681818	0.530303
Missing value method: SKNN-5				
Discretization method: ChiMerge				
Naive Bayes	0.856061	0.916667	0.795455	0.719697
KNN-5	0.840909	0.901515	0.772727	0.681818
C4.5	0.787879	0.924242	0.765152	0.643939
Boosting	0.825758	0.878788	0.825758	0.666667
Bagging	0.833333	0.946970	0.689394	0.651515
SVM	0.765152	0.833333	0.628788	0.568182
Multilayer perceptron	0.840909	0.901515	0.833333	0.659091

TABLE 6

Changes in Classification Accuracy for the Data Set after Removing Both the Features With High Missing Rates and Outliers

Classifier	Type_2_0.3	Type_2_0.5	Type_3	Type_4
Missing value method: Mean imputation				
Discretization method: Bin-5				
Naive Bayes	0.022727	0.151515	0.022727	-0.03025
KNN	-0.10606	-0.03788	-0.11364	-0.06818
C4.5	-0.12879	-0.06818	0.060606	-0.03788
Boosting	-0.01515	-0.01515	0.007579	0.022727
Bagging	-0.01515	-0.08333	0.05303	-0.04546
SVM	-0.0303	0.022728	0.015152	0.022727
Multilayer perceptron	-0.02273	0.022727	-0.0303	-0.02273
Missing value method: Mean imputation				
Discretization method: ChiMerge				
Naive Bayes	0.037882	-0.01515	-0.00758	-0.02273
KNN	-0.01515	0	0.015151	0.007576
C4.5	0.022728	0	0.007576	-0.03788
Boosting	0.030303	0	0.037879	-0.06061
Bagging	0	0	-0.01515	-0.04545
SVM	0.015152	0.015151	-0.0303	-0.02273
Multilayer perceptron	0.030303	-0.00758	-0.00758	-0.02273
Missing value method: SKNN-5				
Discretization method: Bin-5				
Naive Bayes	-0.06818	0	-0.05303	-0.05303
KNN-5	-0.03788	-0.03788	-0.02273	-0.01515
C4.5	-0.06061	0.075757	-0.03788	-0.07576
Boosting	-0.04545	-0.00758	-0.05303	-0.02273
Bagging	-0.03788	-0.00758	0	0.045455
SVM	0.007576	0	0.075757	0.106061
Multilayer perceptron	-0.06818	-0.01515	-0.05303	-0.09091
Missing value method: SKNN-5				
Discretization method: ChiMerge				
Naive Bayes	-0.06818	-0.01515	-0.06818	-0.06061
KNN-5	-0.08333	-0.08333	-0.09091	-0.08333
C4.5	-0.04546	-0.09091	-0.10606	-0.03788
Boosting	-0.09849	0	-0.12121	-0.06818
Bagging	-0.10606	-0.07576	-0.01515	-0.08333
SVM	0.037878	-0.0303	0.030303	0.030303
Multilayer perceptron	-0.02273	-0.03788	-0.09849	0

clear dose-related difference between the low dose and the high and medium doses. There were also drugs and chemicals that did not have major effects on most immunological parameters and did not affect resistance to *Listeria* (Supplementary table S1). Using each of these data cases independently (rather than only comparing them with regard to dose response effects) provided sufficient patterns of response to allow machine learning methods to identify predictive patterns.

Due to the noise and missing data present in the NTP data set, a sequence of data preprocessing steps had to be conducted to improve the data quality, including feature selection, outlier detection, missing data imputation, and variable discretization. We have discussed feature selection and outlier detection in an earlier section. For missing data imputation, we compared the effect of the mean-value imputation and Sparse-KNN methods in Figure 7 using two example attributes. Figure 7a shows the

attribute CD4⁺ cells. About 75% of the values are missing for this attribute. We did not exclude this attribute because it is considered to be involved in an important way in resistance to *L. monocytogenes*. The blue solid line with open circle symbols corresponds to the values after using the mean imputation method, whereas the red line with asterisk symbols correspond to those values of the Sparse-KNN method. The overlapping points of the two lines are the known values in the original data set, while the others are imputed values. Because the mean imputation method fills each missing value with the same value, there is very little variation in the final values of CD4⁺. This makes the attribute ineffective in predicting the class variable. In comparison, the Sparse-KNN method was able to impute more meaningful values by using the mean values of the nearest neighbors of each individual data case. It is also

TABLE 7
Changes in Classification Accuracy for the Data Set after
Removing Only the Features with High Missing Rates (Not
Removing Outliers)

Classifier	Type_2_0.3	Type_2_0.5	Type_3	Type_4
Missing value method: Mean imputation				
Discretization method: Bin-5				
Naive Bayes	0.017805	0.07513	0.023837	0.006517
KNN	-0.08189	-0.00391	-0.03643	0.049845
C4.5	-0.06794	-0.09539	0.077205	0.04227
Boosting	-0.0719	-0.09636	0.021766	0.080148
Bagging	0.023307	-0.08965	0.040532	-0.03745
SVM	-0.11976	0.0152	0.002992	0.052258
Multilayer perceptron	0.016068	-0.02818	-0.05588	-0.03537
Missing value method: Mean imputation				
Discretization method: ChiMerge				
Naive Bayes	-0.00531	-0.12169	0.041015	0.002123
KNN	-0.03387	-0.03749	0.003329	-0.05226
C4.5	0.023451	-0.11687	0.078218	-0.0304
Boosting	0.023789	-0.05834	-0.01631	-0.07156
Bagging	-0.02321	-0.05076	0.013994	-0.02644
SVM	0.029676	-0.0027	-0.00338	0.057421
Multilayer perceptron	-0.00444	-0.02804	0.040339	-0.03884
Missing value method: SKNN-5				
Discretization method: Bin-5				
Naive Bayes	-0.11219	-0.07397	-0.11407	-0.13419
KNN-5	-0.14008	-0.06331	-0.09342	-0.06017
C4.5	-0.10582	-0.0027	-0.09067	-0.00473
Boosting	-0.19345	-0.08	-0.12131	-0.10114
Bagging	-0.10548	-0.01906	-0.07069	-0.04777
SVM	-0.0843	-0.00994	0.048446	0.069823
Multilayer perceptron	-0.15523	-0.05332	-0.19137	-0.1354
Missing value method: SKNN-5				
Discretization method: ChiMerge				
Naive Bayes	-0.02167	-0.00584	0.000723	-0.02543
KNN-5	-0.10843	-0.13082	-0.11668	-0.12768
C4.5	-0.12546	-0.1408	-0.08999	-0.12165
Boosting	-0.05506	-0.03165	-0.05506	-0.06794
Bagging	-0.04352	-0.13168	0.004874	-0.16744
SVM	-0.14095	-0.10085	0.027263	-0.07137
Multilayer perceptron	-0.07021	-0.04164	-0.12633	-0.054

clear from Figure 7b that the values imputed to the attribute Ig⁺ by Sparse-KNN have much more variety than those imputed by the mean imputation. The values obtained using Sparse-KNN were, therefore, more representative of parameters in this study for which there are few missing data cases.

Table 5 shows the accuracy of the classification algorithms in predicting the resistance to infection for the four data sets generated by the combination of two data imputation methods (mean imputation and Sparse-KNN) and two discretization methods (equal-width interval and ChiMerge). Although the results in Figure 7 suggested that Sparse-KNN yielded more realistic imputed results, the key issue is which method has the best predictive performance. Table 5 shows the results when comparing the predicted and known values. All the classifiers had the worst accuracy on the data set produced by mean-value

imputation and equal-width interval discretization. The worst accuracy was as low as around 45% on the Type_4 data set. Using Sparse-KNN instead to impute the missing values immediately improved the accuracy of all the classifiers, and using ChiMerge instead for discretization improved the accuracy results even more. Finally, the classifiers had the best overall performance on the data set produced by using the combination of Sparse-KNN and ChiMerge. The final accuracy results on the Type_2_0.5 data set were as high as 94.7% for many of the classifiers and were around 80% for the Type_2_0.3 and Type_3 data sets. The accuracy results on the Type_4 data set were much lower, however. Because these patterns were similar across all the different classifiers, we can safely conclude that Sparse-KNN is more effective than the simple mean imputation method, and the supervised discretization method ChiMerge performs much better than the unsupervised equal-width discretization method.

Another observation is that, for the same classifier, the accuracy results typically deteriorate as the number of categories in the class variable (mortality) increases. For example, the neural network classifier has an accuracy rate equal to 90% on the Type_2_0.5 data set, but its accuracy rate decreased to 83% for the Type_3 data set and 66% for the Type_4 data set. These results are expected because the more the categories, the harder the classification task. However, the results are similarly useful when 2–4 categories are used. For example, 50% or greater mortality (as in Type_2) is typically indicative of meaningful immunological suppression, whereas less than 50% can be within the range of normal variation in control mice using this *Listeria* model. The other binning protocols with three or four bins allow finer distinction between normal (generally 0–20%) mortality and elevated mortality.

Overall, most of the classifiers have accuracy rates between 85–94% for the Type_2_0.3 (0.3 as the dividing point) and Type_2_0.5 (0.5 as the dividing point) data sets, 75–85% for the Type_3 data set, and lower than 72% for the Type_4 data set. Most classifiers had better accuracy on the Type_2_0.5 data sets produced by the Sparse-KNN data imputation method than on the Type_2_0.3 data sets. This result is somewhat counterintuitive because 0.3 seems a more natural discretization threshold than 0.5 according to expert knowledge. On the other hand, the classifiers typically had worse performance on the Type_2_0.5 data sets produced by the mean-imputation method. The variance in the performance of the classifiers can be partially attributed to the subtle interaction between data imputation and data discretization.

The results in Tables 6 and 7 show that both deleting attributes with high percentages of missing values and retaining outlier values had some effect on classification accuracies. Deleting the variables with a large percentage of missing values slightly decreased the accuracies of some classifiers by several percent, but improved those of some other classifiers slightly. The overall effect was minor, suggesting that the Sparse-KNN method of imputation of missing data was effective. However, Table 7

FIG. 6. Results for selected immunological parameters and resistance to *Listeria monocytogenes*. Values shown are normalized to the control value (which was arbitrarily defined as 100%) for each parameter, except for % mortality after challenge with *Listeria*, which is shown as the actual value. The four bars for each parameter represent the vehicle control and the low, medium, and high dose of the indicated compounds. The immunological parameters plotted are the ones in the NTP data set known or expected to be involved in resistance to *Listeria*.

FIG. 7. Comparison of values imputed by the mean imputation and Sparse-KNN methods for two example attributes: CD4+ and Ig+. The two lines in each graph stand for the two imputation methods (mean and SKNN_5). The parts where the two lines overlap stand for values present in the original data set, whereas the parts where the two lines differ show that different values were imputed by the two methods. The values imputed by Sparse-KNN clearly have much more variety than those by the mean imputation method.

shows that retaining outlier values has a larger impact on the accuracy results for most classifiers. It reduced the accuracies of some classifiers for as much as 10%.

Among the classification methods we applied to the NTP data set, the decision tree method (C4.5) produces easily understandable models. Figure 8 shows decision tree models obtained for the Type_4 data set produced by Sparse-KNN and ChiMerge. A decision tree iteratively branches on the value of the most discriminative attribute conditional on attributes already used in attempting to classify each data case to one of the categories. When a new data case arrives, it is classified using a decision tree as follows. First, the root attribute of the decision tree is checked for its value. If the value leads to a leaf node, the category associated with the node is what the case belongs to. Otherwise, if the value leads to an internal node, the attribute associated with that internal

node is further checked. The above process is repeated until a leaf node (hence, a unique class) is found. In the decision tree of Figure 8, the root attribute is CD4+, a biological factor that is considered to have a large impact in resistance to *L. monocytogenes*, particularly in the model used in this study, in which mortality may occur late (more than 14 days) in the infection (North *et al.*, 1997; Edelson and Unanue, 2000). Depending on the value of CD4+, the next attribute used to branch the decision tree varies. Each path from the root to a leaf is regarded as a decision rule for classifying data cases. The variables closer to the root of the decision tree are considered to be more important in the classification tasks. This method is useful because it not only classifies cases (effect of a particular dose of a particular drug or chemical), but also indicates which immunological parameters were most important in obtaining this classification.

FIG. 8. Decision tree for the Type_4 data set of resistance to *L. monocytogenes* produced using ChiMerge and Sparse-KNN. Immunological parameters are indicated in rectangular boxes. The most discriminative attribute (immunological parameter) with regard to resistance to *Listeria* is at the top and parameters below this have relatively less discriminative power. The square boxes indicate discretization categories for the prediction of mortality due to *L. monocytogenes*: A = 0–0.25; B = 0.25–0.50; C = 0.50–0.75; D = 0.75–1.00. The Type_4 discretization set was used for the classifying variables (immunological parameters). This set includes **four** bins (0–0.25; 0.25–0.5; 0.5–0.75; 0.75–1.0). Some of the variables did not have values in one of the bins, so only two are shown in the figure. A value that falls into the first discretization category ($CD4_cells = 1$) is associated with a different category of mortality than a value that falls into discretization category 2 ($CD4_cells = 2$). Analysis of this type for all doses (and controls) for all drugs and chemicals yielded the structure illustrated here. This method objectively identified most of the parameters in the NTP data set that are known to be important in resistance to *Listeria*.

DISCUSSION

In this paper, we analyzed a portion of the NTP immunotoxicology data set using a variety of machine learning classification methods. The results show that the machine learning methods can accurately predict changes in resistance to *L. monocytogenes* on the basis of changes in 15 immunological parameters with accuracy as high as 94.7%. No other methods have been reported to yield greater accuracy, and the accuracy approaches obtained using commercial software was designed to predict health outcomes in disease states not including infectious diseases (e.g., www.entelos.com).

To be able to do the machine learning analysis, however, several data preprocessing steps had to be performed on the NTP data set. This is due to the significant amount of noise and missing data present in the NTP data set. Because most real-world biological data sets have the same drawbacks as the NTP data set, this study provides a useful guideline/roadmap for the analysis of such data sets. First, feature selection is used to rule out

attributes that are not very relevant to the prediction task, and outlier detection is used to rule out data cases that are likely to contain errors. Determination of features (immunological parameters in this study) to be used was based on deleting one or more features with substantial collinearity with other features, and on deleting features that exhibited dose-responsive effects for few if any chemicals, for which dose-responsive effects were noted with resistance to infection. This simply indicates that under these conditions, these parameters do not contribute in a meaningful way to dose-responsive changes in resistance to infection, so they were excluded. Second, data imputation was used to fill in missing values. In general, we had to treat two kinds of missing values differently. Values that were missing according to specific mechanisms needed to be imputed manually. Other values that are missing at random could be imputed using automatic data imputation methods. We developed a method called Sparse-KNN, an improvement over the standard KNN data imputation method, to handle data sets with missing values distributed in all immunological

parameters and all test agents. Third, continuous variables were discretized before being analyzed with machine learning algorithms. This study shows that supervised discretization methods such as ChiMerge are typically preferable to unsupervised methods such as equal-width discretization.

The results in Tables 5–7 suggest that the method we used for handling missing values did not introduce significant bias, as the changes in predictive accuracy were not significant. However, the results also show that removing outliers is important for the analysis, as they will help us identify and delete values that do not fall within the normal value range of the attributes and improve the performance of almost all classifiers. If most of the immunological parameters had large percentages of missing data, it is unlikely that prediction accuracy would have been maintained. However, this data set included several parameters with few if any missing data cases, and the method described in this study should be applicable to most data sets that have this characteristic.

As stated previously, the most important variables according to the decision tree method include CD4⁺ cells, CTL, SCell, PFC/spleen, and MLR. These parameters include measures of lymphocyte number and function that would be expected to be important in the resistance to *L. monocytogenes* (Edelson and Unanue, 2000; North *et al.*, 1997). In fact, MLR is regarded to be very similar in nature to the T-cell mediated response required to eliminate *Listeria* (Sasaki *et al.*, 1990). Of course, it would be expected that decreased numbers of lymphocytes would also diminish resistance. Although it has previously been thought that immunity to *L. monocytogenes* is mediated almost entirely by cell-mediated mechanisms, recent studies demonstrate a surprisingly important role for antibody (Edelson and Unanue, 2000). Thus, it is interesting that PFC/spleen, which measures specific antibody responses, is a variable selected by decision tree analysis.

It must be acknowledged that the NTP data do not include a parameter that indicates the efficacy of phagocytic cells and inflammation, other than the WBC count. This may very well be the basis for the failure of the machine learning methods to provide an even higher percentage of correct prediction of host resistance. In one case at least, it is highly likely that innate immunity explains the change in host resistance. It is well established that neutrophils are important in the early stages of resistance to *Listeria* (Conlan and North, 1994). One of the compounds tested in the NTP data set was morphine sulfate. Morphine suppresses a number of lymphocyte parameters, but it actually increases the percentage and number of neutrophils in the blood, and it also significantly increases survival of mice treated with *Listeria* (LeVier *et al.*, 1994). Inclusion of a parameter that measured the change in neutrophil percentage and number (differential count) and/or function would probably have provided an independent variable that would have permitted better prediction of this effect. Increased neutrophil percentages are often associated with stress responses, and a number of drugs and chemicals, including morphine, are stressors (Schwab *et al.*, 2005). It is possible that this increase in neutrophils compensates for the loss of other mechanisms of resistance and leads to effective early clearance

of *L. monocytogenes*. Thus, adding innate immune parameters might make the NTP data even more useful.

The results reported in this study have practical implications. Predicting changes in resistance to infection with reasonable accuracy on the basis of easily measurable immunological parameters would decrease the number of animals needed for thorough immunotoxicity evaluations by decreasing the need for host resistance testing, which generally requires larger group sizes than assessment of the immunological parameters. This would allow a step forward toward the goals of reducing and refining animal use to decrease the number required and decrease distress (caused by infection in host resistance testing). Data sets for which comparable immunological parameters have been measured in mice and in humans exposed to the same immunotoxicant could be used with the machine learning algorithms to predict the effects of changes in selected immunological parameters in humans on resistance to infection (Selgrade, 1999). Data for some immunological parameters and for susceptibility to infection are available for a few drugs such as chemotherapeutic agents (Xue *et al.*, 2010), and these results could be compared in future studies with results predicted by the machine learning models developed using the full NTP data set to determine the efficacy of these methods to predict resistance to infection in humans.

Results of the type of machine learning analyses shown in this study will be useful in estimating the effect of mild to moderate inhibition of multiple immunological parameters on resistance to infection in animal models. The fact that approaches for obtaining such estimates have not been reported previously illustrates a critical aspect of immunity is not well understood: interactions and quantitative relationships between immunological parameters and host resistance.

It should be pointed out that *L. monocytogenes* was selected for this study because it is the most commonly used host resistance model in the NTP data set. However, other models were also used, and analyses similar to the one described in this study should be feasible for some of them. It will be of interest to determine the similarities and differences between the variables and the machine learning algorithms that best predict resistance to each microbe or cancer cell type. In future studies, we plan to examine this with the idea that may be possible to identify a method that has near optimal predictive value for resistance to both microbes and cancer cells.

There have been only a few reports describing efforts to quantitatively predict the effects of changes in multiple immunological parameters on host resistance to infection (Keil *et al.*, 2001; Luster *et al.*, 1993). Our previous results (Keil *et al.*, 2001) indicate that logistic regression or multiple regression was not very effective (about 60% accurate) in predicting resistance to B16F10 melanoma cells or Group B *Streptococci* (Keil *et al.*, 2001). In a seminal study by Luster and colleagues, no single immunological parameter was found to skillfully predict resistance to infection (Luster *et al.*, 1993). Some combinations of parameters were concordant with changes in resistance,

but it was not possible to derive regression models using one chemical that would predict the effects of other chemicals on resistance to *L. monocytogenes* (Luster *et al.*, 1993). The results reported in this study represent the first example of the application of modern machine learning algorithms to quantitatively predict changes in resistance to infection on the basis of drug- or chemical-induced changes in immunological parameters.

One of the most interesting findings in this study is the indication that relationships between changes in immunological parameters and changes in resistance to infection are not as intuitively obvious as expected. This can be observed in the representative data shown in Figure 6. In addition, results in the NTP data set (see online supplemental information) include cases in which dose-dependent decreases in parameters known to be important in resistance to *Listeria* (e.g., NK cell function) (North *et al.*, 1997) were not consistently associated with decreased resistance to *Listeria*. Thus, it is not entirely clear which combination of changes in which parameters were discerned by the machine learning algorithms to allow prediction of resistance to *Listeria*. The decision tree method provides some indication of the relative importance of each immunological parameter in predicting resistance to infection, but it only ranks the parameters and it does not quantify differences in importance. In future studies, we plan to address this problem with Bayesian network analysis, which can explicitly capture the relationships between the immunological parameters using expert knowledge or machine learning techniques to construct a framework of relationships (e.g., number of T cells influences number of cytotoxic T cell, which influences resistance to some pathogens). The quantitative aspects of the relationships can then be determined by an iterative machine learning process. The advantage of such analysis is that it not only provides predictive information, it quantitatively determines which parameters or patterns seem to contribute most to resistance.

Simultaneous mild or moderate suppression of several immunological parameters (as noted for BZ in Table 1) was typically associated with decreased host resistance. The machine classification methods probably take additive or even synergistic effects of suppression of multiple parameters into account. The only limiting factor is that some such effects would need to be present in the training data set. It is also possible that substantial suppression of one parameter does not have the expected effect on *Listeria* resistance because other parameters are at least partially redundant. Typical regression methods are not well suited to detect such patterns. Therefore, the results shown in this study have implications with regard to future directions of immunological analyses in the emerging field of systems immunology (Germain *et al.*, 2011; Kozhenkov *et al.*, 2011; Schubert, 2011).

The broader field of systems biology is designed to provide an understanding of the behavior of complex biological systems, ultimately expressing this behavior using algorithms that have explanatory and predictive value (Ideker *et al.*, 2001). It should be understood that these efforts are analogous to similar efforts in physics or other fields in which the development of equations describing the behavior of a system do not merely

express correlations, they represent a concrete expression of the mechanism under investigation. In many cases, the mechanistic role of expression or function of a particular molecule or cell in an immune response is already known, but the quantitative relationships between this function and the outcome in terms of host resistance are not known. The results obtained represent an initial effort to identify algorithms that can predict host resistance outcomes on the basis of immunological parameters already known to be important in this outcome. Till date, most efforts in systems immunology have focused on accumulation of high-throughput data sets and developing database annotation and computational tools to describe networks and pathways. However, few studies have focused on developing algorithms that can be used to predict the effects of perturbation in the immune system on host resistance (Gardy *et al.*, 2009). The results presented indicate that machine learning methods can provide such algorithms, which have been validated using different combinations of training and prediction data sets and which we plan to further validate in subsequent studies with additional chemicals and drugs in the NTP database. We suggest that the results presented demonstrate that machine learning algorithms will be useful in the field of systems immunology. Even though the data used in this study do not include every level of biological organization (signaling, gene expression, protein expression, protein interactions, and protein functions), the use of relatively holistic functional variables gives an indication that incorporating lower levels of organization could be done in a similar manner. This could eventually lead to algorithms that quantitatively explain and predict resistance to infection on the basis of molecular mechanisms necessary for this resistance.

SUPPLEMENTARY DATA

Supplementary data are available online at <http://toxsci.oxfordjournals.org/>.

FUNDING

NIH grant R01ES013708, NSF grants EPS-0903787 and IIS-0953723, and by Mississippi State University.

ACKNOWLEDGMENTS

The authors thank Dr Dori Germolec (NIEHS) for making NTP data available to us in Excel format.

REFERENCES

- Aha, D., and Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning* **6**, 37–66.
- Beskid, G., Eskin, B., Cleeland, R., Siebelist, J., Cappetta, A., Hill, A. D., and Geiger, R. H. (1981). Antibacterial activity of 2',3'-dideoxyadenosine in vivo and in vitro. *Antimicrob. Agents Chemother.* **19**, 424–428.

- Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123–140.
- Burns, L. A., Sikorski, E. E., Saady, J. J., and Munson, A. E. (1991). Evidence for arsenic as the immunosuppressive component of gallium arsenide. *Toxicol. Appl. Pharmacol.* **110**, 157–169.
- Chan, K. M. and Allen, S. A. (2002). Infectious pulmonary complications in lung transplant recipients. *Semin. Respir. Infect.* **17**, 291–302.
- Cohen, S., Frank, E., Doyle, W. J., Skoner, D. P., Rabin, B. S., and Gwaltney, J. M. Jr. (1998). Types of stressors that increase susceptibility to the common cold in healthy adults. *Health Psychol.* **17**, 214–223.
- Conlan, J. W. and North, R. J. (1994). *Listeria monocytogenes*, but not *Salmonella typhimurium*, elicits a CD18-independent mechanism of neutrophil extravasation into the murine peritoneal cavity. *Infect. Immun.* **62**, 2702–2706.
- Edelson, B. T. and Unanue, E. R. (2000). Immunity to *Listeria* infection. *Curr. Opin. Immunol.* **12**, 425–431.
- Fayyad, U. M., and Irani, K. B. (1993). Multi-interval discretization of continuous valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 1022–1027. Morgan Kaufmann, Chamberg, France.
- Freier, D. O., and Fuchs, B. A. (1994). A mechanism of action for morphine-induced immunosuppression: Corticosterone mediates morphine-induced suppression of natural killer cell activity. *J. Pharmacol. Exp. Ther.* **270**, 1127–1133.
- Friedman, E. M. and Lawrence, D. A. (2002). Environmental stress mediates changes in neuroimmunological interactions. *Toxicol. Sci.* **67**, 4–10.
- Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Ann. Statist.* **28**, 337–407.
- Fung, J. J., Jain, A., Kwak, E. J., Kusne, S., Dvorchik, I., and Egtesad, B. (2001). De novo malignancies after liver transplantation: A major cause of late death. *Liver Transpl.* **7**(Suppl 1), S109–S118.
- Gardy, J. L., Lynn, D. J., Brinkman, F. S., and Hancock, R. E. (2009). Enabling a systems biology approach to immunology: Focus on innate immunity. *Trends Immunol.* **30**, 249–262.
- Germain, R. N., Meier-Schellersheim, M., Nita-Lazar, A., and Fraser, I. D. (2011). Systems biology in immunology: A computational modeling perspective. *Annu. Rev. Immunol.* **29**, 527–585.
- Glover, M., and Pruett, S. B. (2009). The role of stress mediators in modulation of cytokine production by ethanol. *Toxicol. Appl. Pharmacol.* **239**, 98–105.
- Guo, T. L., McCay, J. A., Brown, R. D., Musgrove, D. L., Butterworth, L., Munson, A. E., Germolec, D. R., and White, K. L. Jr. (2000). Glycidol modulation of the immune responses in female B6C3F1 mice. *Drug Chem. Toxicol.* **23**, 433–457.
- Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations* **11**, 10–18.
- Han, Y. C., Lin, T.-L. and Pruett, S. B. (1993). Thymic atrophy caused by ethanol in a mouse model for binge drinking: Involvement of endogenous glucocorticoids. *Toxicol. Appl. Pharmacol.* **123**, 16–25.
- Han, Y. C., and Pruett, S. B. (1995). Mechanisms of ethanol-induced suppression of a primary antibody response in a mouse model for binge drinking. *J. Pharmacol. Exp. Ther.* **275**, 950–957.
- Herbert, T. B. and Cohen, S. (1993). Stress and immunity in humans: A meta-analytic review. *Psychosom. Med.* **55**, 364–379.
- Holsapple, M. P., White, K. L. J., McCay, J. A., Bradley, S. G., and Munson, A. E. (1988). An immunotoxicological evaluation of 4,4'-thiobis-(6-*t*-butyl-*m*-cresol) in female B6C3F1 mice. 2. Humoral and cell-mediated immunity, macrophage function, and host resistance. *Fundam. Appl. Toxicol.* **10**, 701–716.
- Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372.
- Karrow, N. A., McCay, J. A., Brown, R. D., Musgrove, D. L., Pettit, D. A., Munson, A. E., Germolec, D. R., and White, K. L. (2000). Thalidomide stimulates splenic IgM antibody response and cytotoxic T lymphocyte activity and alters leukocyte subpopulation numbers in female B6C3F1 mice. *Toxicol. Appl. Pharmacol.* **165**, 237–244.
- Karrow, N. A., Guo, T. L., Zhang, L. X., McCay, J. A., Musgrove, D. L., Peachee, V. L., Germolec, D. R., and White, K. L. (2003). Thalidomide modulation of the immune response in female B6C3F1 mice: A host resistance study. *Int. Immunopharmacol.* **3**, 1447–1456.
- Keil, D., Luebke, R. W., and Pruett, S. B. (2001). Quantifying the relationship between multiple immunological parameters and host resistance: Probing the limits of reductionism. *J. Immunol.* **167**, 4543–4552.
- Kerber, R. (1992). Chimerge: Discretization of numeric attributes. In *Proceedings Ninth National Conference on Artificial Intelligence*, pp. 123–128. MIT Press, San Jose, CA.
- Kozhenkov, S., Sedova, M., Dubinina, Y., Gupta, A., Ray, A., Ponomarenko, J., and Baitaluk, M. (2011). BiologicalNetworks—tools enabling the integration of multi-scale data for the host-pathogen studies. *BMC Syst. Biol.* **5**, 7.
- LeVier, D. G., McCay, J. A., Stern, M. L., Harris, L. S., Page, D., Brown, R. D., Musgrove, D. L., Butterworth, L. F., White, K. L., and Munson, A. E. (1994). Immunotoxicological profile of morphine sulfate in B6C3F1 female mice. *Fundam. Appl. Toxicol.* **22**, 525–542.
- Luster, M. I., Munson, A. E., Thomas, P. T., Holsapple, M. P., Fenters, J. D., White, K. L., Lauer, L. D., Germolec, D. R., Rosenthal, G. J., and Dean, J. H. (1988). Methods evaluation. Development of a testing battery to assess chemical-induced immunotoxicity: National Toxicology Program's guidelines for immunotoxicity evaluation in mice. *Fundam. Appl. Toxicol.* **10**, 2–19.
- Luster, M. I., Portier, C., Pait, D. G., White, K. L. Jr, Gennings, C., Munson, A. E., and Rosenthal, G. J. (1992). Risk assessment in immunotoxicology. I. Sensitivity and predictability of immune tests. *Fundam. Appl. Toxicol.* **18**, 200–210.
- Luster, M. I., Portier, C., Pait, D. G., Rosenthal, G. J., Germolec, D. R., Corsini, E., Blaylock, B. L., Pollock, P., Kouchi, Y., Craig, W., et al. (1993). Risk assessment in immunotoxicology. II. Relationships between immune and host resistance tests. *Fundam. Appl. Toxicol.* **21**, 71–82.
- McCallum, A. and Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, pp. 41–48.
- Nathwani, A. C., Davidoff, A. M., and Linch, D. C. (2005). A review of gene therapy for haematological disorders. *Br. J. Haematol.* **128**, 3–17.
- North, R. J., Dunn, P. L., and Conlan, J. W. (1997). Murine listeriosis as a model of antimicrobial defense. *Immunol. Rev.* **158**, 27–36.
- Pal, S. K., and Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. *IEEE Trans. Neural Netw.* **3**, 683–697.
- Pruett, S. B., Schwab, C., Zheng, Q., and Fan, R. (2004). Suppression of innate immunity by acute ethanol administration: A global perspective and a new mechanism beginning with inhibition of signaling through TLR3. *J. Immunol.* **173**, 2715–2724.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, CA.
- Sasaki, T., Mieno, M., Udono, H., Yamaguchi, K., Usui, T., Hara, K., Shiku, H., and Nakayama, E. (1990). Roles of CD4+ and CD8+ cells, and the effect of administration of recombinant murine interferon gamma in listerial infection. *J. Exp. Med.* **171**, 1141–1154.
- Schubert, C. (2011). Systems immunology: Complexity captured. *Nature* **473**, 113–114.
- Schwab, C. L., Fan, R., Zheng, Q., Myers, L. P., Hébert, P., and Pruett, S. B. (2005). Modeling and predicting stress-induced immunosuppression in mice using blood parameters. *Toxicol. Sci.* **83**, 101–113.
- Selgrade, M. K. (1999). Use of immunotoxicity data in health risk assessments: Uncertainties and research to improve the process. *Toxicology* **133**, 59–72.

- Sikorski, E. E., McCay, J. A., White, K. L. J., Bradley, S. G., and Munson, A. E. (1989). Immunotoxicity of the semiconductor gallium arsenide in female B6C3F1 mice. *Fundam. Appl. Toxicol.* **13**, 843–858.
- Tufano, R. (2002). Focus on risk factors for fungal infections in ICU patients. *Minerva Anesthesiol.* **68**, 269–272.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.
- Vedhara, K., Cox, N. K., Wilcock, G. K., Perks, P., Hunt, M., Anderson, S., Lightman, S. L., and Shanks, N. M. (1999). Chronic stress in elderly carers of dementia patients and antibody response to influenza vaccination. *Lancet* **353**, 627–631.
- Xue, F., Zhang, J., Han, L., Li, Q., Xu, N., Zhou, T., Xi, Z., Wu, Y., and Xia, Q. (2010). Immune cell functional assay in monitoring of adult liver transplantation recipients with infection. *Transplantation* **89**, 620–626.
- Zandman-Goddard, G. and Shoenfeld, Y. (2003). SLE and infections. *Clin. Rev. Allergy Immunol.* **25**, 29–40.