

## Validation of Epidemiological Models: Chicken Epidemiology in the UK

Dmitriy Fradkin, Ilya Muchnik, Patrick Hermans, and Kenton Morgan

**ABSTRACT.** In epidemiology, a standard way of constructing models is to conduct univariate analysis of independent variables, followed by fitting a multivariate logistic model to the selected features. The main measure for choosing a particular model is a goodness of fit criterion on a given dataset. While this measure indicates how well the model fits the data, it has little relation to the predictive accuracy of the model and therefore may not generalize beyond the given dataset. This aspect is not frequently considered in epidemiology. We suggest using modern machine learning methods for constructing and validating epidemiological models. The resulting models can be used to confirm epidemiologist's models or to suggest possible improvements. These approaches also provide estimates and confidence measures for the parameters and the predictive ability of the model.

### 1. Overview

In epidemiology, a standard way of constructing models is to conduct univariate analysis of independent variables, followed by manual selection and refinement of the feature set by fitting a multivariate logistic model to the data [3]. The model constructed on the final set of variables is seen as the final model. The main measure driving the model selection process is therefore a goodness of fit criterion on a given dataset. While this measure indicates how well the model fits the data, it has little relation to the predictive accuracy of the model and therefore may not generalize beyond the given dataset. However, from the point of view of prevention, it is extremely important to identify the factors responsible for the presence or absence of a disease or a symptom on data that has not yet been observed.

Model validation (estimation of predictive performance) is not frequently done in human epidemiology and medicine, as noted for example in [11, 14], despite being recognized as an important part of establishing the usefulness of a model. Validation is even more rare in the field of veterinary epidemiology. A few exceptions ([5, 12]) describe models validated using hold-out test sets.

---

2000 *Mathematics Subject Classification.* 62H30,68T05.

*Key words and phrases.* support vector machines, feature selection, cross validation.

Dr. Muchnik's work was supported through NSF grant CCR 0325398.

Dr. Hermans and Dr. Morgan acknowledge Elanco Animal Health and the Engineering and Physical Sciences Research Council (EPSRC) that funded their work.

The main goal of our work is to demonstrate how modern machine learning methods can be used to conduct statistical validation of epidemiological models in the absence of an independent test set. We found one publication addressing this issue, [9], but the leave-one-out method used there was applied for spectroscopic data, where the measurements are much more accurate than in the standard epidemiological cross-sectional data. It is much harder to have validation based on statistical estimates when one uses noisy cross-sectional data. This paper focuses specifically on such situations.

The principal tool for this task is the cross-validation technique, used in machine learning for obtaining estimates of performance and evaluation of parameters [10]. Additionally, we examine a recently proposed feature selection method that was shown to be effective in a number of applications [6]. In contrast to univariate feature selection, which is standard in epidemiology, it is a multivariable selection method. This means that it is more general: it will work in the cases where univariate selection works, but it could also work in cases where univariate selection does not. While cross-validation and feature selection methods are not new, they are little known to epidemiologists. We believe that their usage would lead to better models.

We illustrate our suggestions on a dataset describing the occurrences of wet litter in the poultry farms in the UK in the year 2001, collected and investigated using traditional methods by Hermans et. al. [8]. The methods discussed, however, can be applied in validation of different classification-based epidemiological models.

We want to emphasize that the logistic models can also be interpreted as classifiers because the data itself is labeled into classes; moreover, frequently only two classes (positive/negative; disease/no disease, etc.) occur. When one estimates a fitness score of the constructed model, the number of errors given by the model on the data should be calculated rather than the quality of the fit.<sup>1</sup>

## 2. Background

**2.1. Wet Litter.** Here we present a brief introduction to the significance and causes of wet litter. For a detailed discussion a reader is referred to [8].

Wet litter is a term used when the material covering the floors of poultry houses, usually consisting of wood shavings or chopped straw, reaches its saturation threshold and is incapable of holding more moisture. Wet litter prevention and control is important for the health, welfare and productivity of broiler flocks. It has been shown to be associated with the occurrence of foot-pad, breast and hock lesions, and inducing high levels of stress in the birds. A number of risk factors have been identified in the literature.

The analysis of longitudinal data collected from 639 farms (75% of those where the questionnaire was sent) showed that more than half the farmers had wet litter in

---

<sup>1</sup>Strictly speaking, a logistic model assigns to an observation the (estimate of) probability that the observation belongs to one particular class. In practice this estimate is compared with the threshold 0.5. If the estimate is larger than this threshold an instance is assigned to one class, and, if not, then to the other. The probability is given by formula:

$$(1.1) \quad P(\text{class } 1|x) = \frac{\exp(w \cdot x + b)}{1 + \exp(w \cdot x + b)}$$

where coefficients  $w$  and  $b$  determine the model [7]. It is very important to note that the formula does not give a simple way to measure significance for different variables.

Symbol	Meaning
$d$	dimensionality of the data
$x, y \in X$	data points, belonging to a space $X$
$x \cdot y$	inner product between two vectors
$\ x\ $	$L_2$ norm of vector $x$
$w$	a weight vector
$b$	bias

TABLE 1. Notation

their most recent flock. This indicates that wet litter is a wide-spread phenomenon and that better understanding of its causes is needed to reduce costs to the poultry industry.

**2.2. Notation.** Before proceeding further we would like to specify our notation. Symbols are described in Table 1. Throughout this paper we use the words “vector” and “point” interchangeably, depending on which is more suitable in the context of the discussion. Both terms indicate a group of labeled factors. Below we describe instances only by numerical factors. Taken together, these factors form a Euclidean space (the space of numerical variables). The number of factors determines the dimensionality of the space.<sup>2</sup>

**2.3. Cross-Validation.** The theory of cross-validation [4] is based on two ideas:

- Predictions of the model on new data give the strongest statistical estimates of the model validity.
- There exist re-sampling schemes of the restricted data that can be used to validate a model without use of a hold-out set.

One of the most popular re-sampling schemes is  $k$ -fold cross-validation [10]. This procedure consists of  $k$  identical stages as follows. The set of observations is divided into  $k$  equal subsets. Let us choose one of these to be the “test” set. The remaining  $k - 1$  subsets are then combined to form a “training” set. The training set is used to build a model that is evaluated on the test set. In the next stage, a different subset is chosen to be a test set, and the remaining subsets are combined into a training set. This is repeated  $k$  times, using each of the  $k$  subsets in turn as the test set. Notice that as a result we obtain a prediction for each available observation, allowing us to estimate the predictive performance of the model built on the whole data, and  $k$  estimates of model coefficients. If this procedure is repeated  $t$  times (with a different partition into  $k$  sets every time), we have  $kt$  estimates on the coefficients, and  $t$  estimates of the predictive accuracy of the model on all observations.

Note that the cross-validation method estimates an average performance from an ensemble of models and proposes to use them as estimates of performance of a single model identified on the whole set of observations. A proof of the correctness of this approach can be found in [4].

This test has several advantages for validation of complex statistical models:

---

<sup>2</sup>In practice many factors originally are not numerical, but we transform the original heterogeneous factors in order to be able to represent data as points in a uniform Euclidean space. This type of space has properties which are important for constructing models.

- It is independent of the type of distribution function, and therefore its conclusions about model validity are most rigorous.
- The test directly measures variance of all the coefficients of the model; this is very convenient from the practical point of view.
- If the test confirms that a model has a high validity, the coefficients can be used to numerically compare the significance of variables. Moreover, because the validity is also measured numerically, it is possible to compute confidence of the model validity estimates.

A stratified cross-validation schema differs from regular cross-validation in that the test set is required to have the same distribution of classes as the whole set. Another way to think of this is that each class is partitioned into folds separately and the corresponding folds are then combined.

A number of papers examined the quality of estimates produced by cross-validation. In the experiments reported in [10], the k-fold cross-validation was pessimistically biased (that is, true accuracy was higher than the estimate) for  $k = 2, 5$ , although the estimates improved for  $k=10, 20$ . Stratified cross-validation behaved similarly, but with a lower bias.

So the recommendation of [10] is to use stratified 10-fold cross-validation. In this work we follow that advice.

## 2.4. Linear Classifiers.

2.4.1. *Classifier Constructed by Logistic Regression.* We have mentioned in the Section 1 that in the epidemiological practice the logistic model is used as a classifier. A standard way to transform logistic regression into a classification model is to consider the logarithm of the ratio of probabilities that an observation belongs to one or the other class. Using formula (1.1), it is easy to show that this approach results in a linear classifier  $f(x)$ , given by:

$$(2.1) \quad f(x) = \begin{cases} 1 & \text{if } w \cdot x + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

where  $\{-1, +1\}$  are the class labels. This view of has an important advantage over using the original non-linear function of equation (1.1) to calculate class probabilities. According to formula (2.1), the vector of coefficients,  $w$ , can be directly interpreted as measuring the relative significance of the factors. For instance, the absolute values of the coefficients,  $|w_i|$ , can be used as significance coefficients.

2.4.2. *Optimal (SVM) Classifier.* A linear classifier can be constructed in ways other than by fitting a logistic model. There are many different methods for doing this [7]. In this section we will describe the so-called “support vector machine” (SVM) classifier. The SVM classifier possesses an optimality property: the weight vector  $w$  found by this method is perpendicular to the maximal margin between the classes, and this guarantees, in a statistical sense, a low error rate on new observations. These mathematical results are “distribution-free”, that is, they do not rely on any assumptions about the distribution of the data. It is the presence of strong and general theoretical foundation, supported by empirical evidence of excellent performance on many applications, that makes SVM such a popular machine learning tool. A detailed discussion of SVM methodology can be found in [2, 13]. A short description is contained in the introductory chapter of this volume.

We have to emphasize a particular feature of SVM that is important for practitioners, and particularly, we believe, for epidemiologists, who work with noisy

data, i.e., with data containing a lot of missing values, with subjective responses and different scales and types of factors. In the process of constructing a classifier, the SVM method partitions the data into two sets, corresponding to “important” and “not important” observations. The important ones are exactly the so-called “support vectors”. They form a boundary for the class margin region. In other words, the support vectors are points from each class that are most similar to the points of the other class. For epidemiologists the “not important” observations should provide prototypical examples of observations and feature ranges in each class.

**2.5. SVM-based Feature Selection.** This approach [6] is based on the observation that the feature  $i$  with the smallest absolute value of weight  $|w_i|$  is the one that has the least influence on the decision. This suggests the procedure described in Algorithm 1 for recursive feature elimination (RFE). At each step a feature with the smallest weight is removed from all instances, and SVM is retrained on the remaining features. The order in which the variables are removed can be seen as a ranking of the features. Such a ranking can be used in two ways:

- (1) It provides means of feature reduction by keeping the most influential features. Of course the phrase “most influential” is not formal because the heuristic procedure does not guarantee finding the subset of features that would result in the best performance. However experiments show that the features selected by this method are important [6].
- (2) It allows one to interpret the relative influences of the features on the classification. From this perspective, it is interesting to apply Algorithm 1 to ranking the whole set of features.

---

**Algorithm 1** SVM Recursive Feature Elimination (RFE) [6]

---

**Require:** A set of points  $W$  in  $\mathfrak{R}^d$ ;  $f$  - number of desired features

- 1: **for**  $i = d, \dots, f + 1$  **do**
  - 2:   Construct an SVM on set  $W$ . Let  $w^i$  be the decision hyperplane of that SVM.
  - 3:   Let  $k_i = \operatorname{argmin}_j |w_j^i|$ .
  - 4:   Remove feature  $k_i$  from all points  $x$  in  $W$ .
  - 5: **end for**
  - 6: Output resulting classifier with  $f$  features and hyperplane  $w^f$ .
- 

### 3. Experimental Results of Cross-Validation

The risk factors have been selected manually using the univariate and multivariate analysis described above. We repeated 10-fold stratified cross-validation of each wet litter model 10 times. In our experiments, we used the same 444 observations that were used by epidemiologist to construct final model M1 (248 cases with wet litter and 196 cases without it). We used SVM package called LIBSVM [1]. Cross-validation was implemented with a shell script.

Table 2 shows the final model constructed by multivariate regression on the set of features M1. Table 3 show the model constructed by SVM on the same set of features. Only two coefficients (of factors “side ventilation” and quadratic polynomial in number of people) have the coefficient of variation (ratio of standard

Variable	Coefficient	S.E.	O.R.	95% C.I.	p-value
Coccidiosis	1.620	0.654	5.05	1.40-18.19	0.01
Specific breed <sup>4</sup>	-0.794	0.247	0.43	0.27-0.71	0.001
Flock was thinned	1.350	0.402	3.86	1.75-8.49	0.001
Age at slaughter	-0.092	0.022	0.91	0.87-0.95	< 0.001
Side ventilation	0.552	0.236	1.74	1.09-2.76	0.02
Separate farm clothing for each house	-1.116	0.272	0.33	0.19-0.56	< 0.001
Plastic overshoes for each house	0.747	0.235	2.11	1.33-3.34	0.001
Number of rodent baits	0.021	0.007	1.02	1.01-1.04	0.005
Pigs on the farm	1.424	0.665	4.15	1.13-15.29	0.03
Feeding equipment failures	0.701	0.236	2.02	1.27-3.20	0.003
Number of farm workers (quadratic polynomial)	6.162	2.494	474.52	3.57-63012.92	0.01
Number of farm workers (cubic polynomial)	-4.118	1.677	0.016	0.00-0.44	0.01
Constant	2.787	1.019	-	-	-

TABLE 2. Final Epidemiological Model (M1) [8].

deviation to the mean value) greater than 0.2.<sup>3</sup>The stability of model accuracy estimates is very high.

We would like to make some general observations. One is that the standard deviations of the coefficients and of the predictive accuracy measures are small, indicating high reliability of these estimates. Another observation is that the relative magnitudes and signs of the coefficients match those obtained by fitting multivariate logistic regression to all observations. This observation lends support to the interpretation of the importance of different risk factors. Finally, the estimates of predictive accuracy of the model for wet litter are high, indicating that the factors included in the model are indeed significant, and that the model can be used as a basis for making predictions and recommendations regarding wet litter.

#### 4. Choice of Features

In the previous section we discussed how the quality of a model can be measured with cross-validation. Here we discuss how the choice of features can be compared against other possibilities.

We propose comparing the final features of epidemiological model M1, constructed with traditional methods, against the feature set of the same size (M2) chosen completely by RFE method (Algorithm 1). We also consider the effect of using decreased or increased sets of features (M3 and M4 respectively).

In comparing M1 and M2 (or M3), the important aspects are the overlap in the number of features, and the estimates of predictive accuracy. In comparing M1 and M4, the number of features that are “added” into the final set is significant. In both cases, the relative sizes of the coefficients and stability of the results are important.

<sup>3</sup>The coefficient of variation is used in statistics to measure fluctuations of a random variable around the variable’s mean value.

<sup>4</sup>The name of the breed will be kept anonymous until results are released to the industry.

Features	$\mu$	$\sigma$	$ \frac{\sigma}{\mu} $
Coccidiosis	1.074	0.144	0.134
Specific breed	-0.760	0.136	0.179
Flock was thinned	1.196	0.132	0.110
Age at slaughter	-0.076	0.010	0.137
Side ventilation	0.631	0.144	0.229
Separate farm clothing for each house	-1.120	0.086	0.077
Plastic overshoes for each house	0.796	0.097	0.122
Number of rodent baits	0.023	0.003	0.113
Pigs on the farm	1.106	0.133	0.121
Feeding equipment failures	0.756	0.099	0.131
Number of farm workers (quadratic polynomial)	1.134	0.259	0.228
Number of farm workers (cubic polynomial)	-0.879	0.151	0.172
Model Evaluation			
Sensitivity	0.722	0.007	0.010
Specificity	0.620	0.013	0.020

TABLE 3. Cross-validation of M1 Using SVM.

Rank	Features	Model size before removal
12	Number of farm workers (quadratic polynomial)	153
11	Coccidiosis	147
10	Side ventilation	18
9	Pigs on the farm	16
8	Number of rodent baits	15
7	Specific breed	14
6	Plastic overshoes for each house	8
5	Feeding equipment failures	6
4	Age at slaughter	5
3	Number of farm workers (cubic polynomial)	3
2	Flock was thinned	2
1	Separate farm clothing for each house	1

TABLE 4. RFE Ranking of Features of M1.

**4.1. Experimental Results.** The full feature set (after removing features with more than 10% missing values and “constant” features) is of size 177. Many features still have missing values.

In order to avoid convergence problems with SVM we applied statistical normalization to all features. Each feature is independently transformed to have zero mean and unit variance:

$$(4.1) \quad x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

where  $\bar{x}_j$  is the mean value of the  $j$ th coordinate and  $\sigma_j$  is its standard deviation, estimated on the instances where the feature value was present. The missing values were set to 0 (i.e. to the mean feature value).

Features	$\mu$	$\sigma$	Rescaled $\mu$	Rescaled $\sigma$	$ \frac{\sigma}{\mu} $
Coccidiosis	0.299	0.053	1.251	0.223	0.178
Specific breed	-0.382	0.065	-0.768	0.131	0.171
Flock was thinned	0.384	0.043	1.298	0.147	0.113
Age at slaughter	-0.425	0.059	-0.076	0.011	0.139
Side ventilation	0.320	0.071	0.660	0.146	0.221
Separate farm clothing for each house	-0.554	0.040	-1.179	0.084	0.071
Plastic overshoes for each house	0.411	0.048	0.862	0.100	0.116
Number of rodent baits	0.391	0.055	0.022	0.003	0.140
Pigs on the farm	0.230	0.029	1.273	0.159	0.125
Feeding equipment failures	0.378	0.049	0.776	0.100	0.129
Number of farm workers (quadratic polynomial)	0.662	0.205	1.192	0.369	0.309
Number of farm workers (cubic polynomial)	-1.431	0.298	-0.882	0.184	0.209
Model Evaluation					
Sensitivity			0.724	0.011	0.015
Specificity			0.629	0.010	0.015

TABLE 5. Normalized cross-validation of M1.

Features	$\mu$	$\sigma$	Rescaled $\mu$	Rescaled $\sigma$	$ \frac{\sigma}{\mu} $
Wall of concrete	0.239	0.032	0.524	0.070	0.134
Number of farm workers	0.521	0.081	0.230	0.036	0.155
Flock was thinned	0.504	0.071	1.704	0.241	0.141
Age at slaughter	-0.470	0.033	-0.084	0.006	0.069
Who cleaned the houses before last crop	-0.322	0.083	-0.387	0.099	0.257
Who disinfected houses before last crop	0.333	0.080	0.385	0.093	0.241
Separate farm clothing for each house	-0.567	0.029	-1.207	0.062	0.051
Plastic overshoes for each house	0.362	0.038	0.759	0.081	0.106
Broiler farm nearby	-0.203	0.037	-0.406	0.073	0.180
Feeding equipment failures	0.396	0.029	0.813	0.061	0.074
Water supply: borehole	-0.347	0.032	-0.790	0.073	0.093
Number of farm workers (cubic polynomial)	-1.407	0.159	-0.867	0.098	0.113
Model Evaluation					
Sensitivity			0.753	0.009	0.011
Specificity			0.525	0.011	0.021

TABLE 6. Normalized cross-validation of M2 - top 12 RFE features.

Since the normalization is “reversible” (i.e. one-to-one), we give the coefficient means and variances for normalized features, and also show their values if scaled back to the original magnitude. Suppose  $w'_i$  is the coefficient of feature  $i$  after

Features	$\mu$	$\sigma$	Rescaled $\mu$	Rescaled $\sigma$	$ \frac{\sigma}{\mu} $
Number of farm workers	0.191	0.187	0.084	0.083	0.979
Flock was thinned	0.546	0.119	1.846	0.401	0.217
Age at slaughter	-0.085	0.097	-0.015	0.017	1.145
Separate farm clothing for each house	-0.866	0.090	-1.843	0.192	0.104
Number of farm workers (cubic polynomial)	-0.665	0.387	-0.410	0.238	0.582
Model Evaluation					
Sensitivity			0.679	0.004	0.006
Specificity			0.507	0.012	0.023

TABLE 7. Normalized cross-validation of M3 - top 5 RFE features.

normalization. Since

$$\begin{aligned}
 w' \cdot x' + b &= \sum_{i=1}^d w'_i x'_i + b = \sum_{i=1}^d w'_i \frac{x_i - \bar{x}_i}{\sigma_i} + b \\
 &= \sum_{i=1}^d \frac{w'_i x_i}{\sigma_i} + \left( b - \sum_{i=1}^d \frac{w'_i \bar{x}_i}{\sigma_i} \right)
 \end{aligned}$$

and the term in parenthesis in the last line is constant, it follows that coefficient  $w_i$  for unnormalized  $i$ th feature is  $w_i = w'_i / \sigma_i$ .

Note that the rescaled means and coefficients in Table 5, which shows results with the features of the final model (M1), are very similar to coefficients in the Table 3. The normalized solution differs only very slightly from the unnormalized solution.

We implemented the original feature selection algorithm of [6]. As described in Section 2.5, the algorithm removes the least significant features one by one. Therefore, the later the feature is removed, the more important it is for classification. Looking at it from the other direction, the fewer features are left in the model when a given feature is removed, the more important that feature is. Table 4 gives the size of the model each time one of the features from M1 is removed. The most important feature according to this table is ‘‘Separate farm clothing for each house’’, and the least important one is quadratic polynomial of the number of farm workers.

The main comparison is between features of M1 and the top 12 features selected by RFE (M2). Results for the top 12 features (M2), and also for the top 5 (M3) and the top 18 features (M4) are in Tables 6, 7 and 8 respectively.

M1 and M2 have 6 out 12 (in each) features in common. It may seem that combining these features (for total of 18) should give better performance than either one separately. However, we found that this results in sensitivity and specificity of 0.758 and 0.607 respectively and thus is slightly better than M1 in sensitivity and somewhat worse in specificity. (These results are slightly better than M2 in sensitivity, and much better in specificity). We suspect that the features that are not common between M1 and M2 play roles of ‘‘duplicates’’ to each other. Table 9 displays correlation matrix of these 18 features and finds that some of them correlate strongly.

We also experiment with the model built on the 5 top features (M3) and the top 18 features (M4). These numbers were chosen because they correspond to regions of

Features	$\mu$	$\sigma$	Rescaled $\mu$	Rescaled $\sigma$	$ \frac{\sigma}{\mu} $
Wall of concrete	0.330	0.037	0.723	0.080	0.111
Number of farm workers	0.506	0.108	0.224	0.048	0.213
Specific Breed	-0.340	0.052	-0.684	0.105	0.154
Chickens separated by sex	-0.293	0.048	-0.639	0.104	0.163
Flock was thinned	0.443	0.045	1.498	0.151	0.101
Age at slaughter	-0.347	0.052	-0.062	0.009	0.149
Side ventilation	0.255	0.052	0.526	0.106	0.202
Who cleaned the houses before last crop	-0.336	0.090	-0.404	0.108	0.268
Who disinfected houses before last crop	0.250	0.083	0.289	0.096	0.333
Separate farm clothing for each house	-0.568	0.048	-1.209	0.102	0.084
Plastic overshoes for each house	0.425	0.054	0.892	0.114	0.128
Number of rodent baits	0.370	0.042	0.021	0.002	0.112
Pigs on the farm	0.278	0.044	1.539	0.243	0.158
Other poultry farms nearby	0.305	0.064	0.632	0.133	0.210
Broiler farm nearby	-0.316	0.050	-0.632	0.100	0.159
Feeding equipment failures	0.333	0.045	0.684	0.092	0.135
Water supply: borehole	-0.353	0.034	-0.804	0.076	0.095
Number of farm workers (cubic polynomial)	-1.349	0.197	-0.831	0.121	0.146
Model Evaluation					
Sensitivity			0.771	0.010	0.013
Specificity			0.594	0.005	0.009

TABLE 8. Normalized cross-validation of M4 - top 18 RFE features.

overlap between features of M1 and the feature ordering produced by RFE. Table 4 demonstrates that the top 18 SVM features include 10 of the 12 M1 features, and considers the other 2 very insignificant. The 5-feature model was chosen because out of these 5 features only 1 does not appear in M1. Therefore models M3 and M4 in some sense “surround” M1. Evaluation of their quality may suggest possible improvements to M1.

Model M3 has 4 out of 5 features in common with M1. Its sensitivity is only somewhat worse than that of M1 (specificity is noticeably worse), but it has only half the features. Model M4 on the other hand has 18 features, and includes 10 top-ranked (by RFE) features of M1. However its results are only slightly better than those of M2, and are comparable (slightly better Sensitivity and slightly worse Specificity) to those of M1.

We note that when using all features, the sensitivity and specificity averages are 0.599 and 0.515 respectively. Therefore, all three selected feature sets perform much better than the whole feature set.

These experiments suggest that M1 is a very good set of features that balances off a reasonable model size with a good predictive performance. On the other hand, our results suggest that RFE method is capable of finding interesting features, many of which match those selected by epidemiologist. It may be interesting to examine some additional features in M4 and to consider effects of possible strong correlations

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	—	32	12	05	06	02	-01	-13	-07	09	06	03	04	04	09	06	06	04
2	32	—	-07	05	08	02	-02	01	-08	09	05	01	08	03	01	-05	-05	06
3	12	-07	—	16	-10	-03	-04	-32	-18	-26	16	-02	09	-02	25	30	-04	08
4	05	05	16	—	-02	04	01	-01	04	07	03	04	-00	05	-02	-05	03	-08
5	06	08	-10	-02	—	08	09	09	04	24	-05	10	-09	14	-20	-22	04	-20
6	02	02	-03	04	08	—	-02	05	06	06	-01	97	-04	79	-05	-05	05	-02
7	-01	-02	-04	01	09	-02	—	-06	09	04	06	-02	06	-02	-12	-11	-05	-09
8	-13	01	-32	-01	09	05	-06	—	18	22	-07	05	-01	09	-24	-24	08	07
9	-07	-08	-18	04	04	06	09	18	—	10	-09	05	06	05	-15	-13	03	04
10	09	09	-26	07	24	06	04	22	10	—	-07	08	-16	14	-17	-17	01	-08
11	06	05	16	03	-05	-01	06	-07	-09	-07	—	-02	02	-04	06	06	-06	00
12	03	01	-02	04	10	97	-02	05	05	08	-02	—	-05	90	-07	-08	04	-01
13	04	08	09	-00	-09	-04	06	-01	06	-16	02	-05	—	-06	01	01	-00	15
14	04	03	-02	05	14	79	-02	09	05	14	-04	90	-06	—	-11	-13	02	02
15	09	01	25	-02	-20	-05	-12	-24	-15	-17	06	-07	01	-11	—	81	-03	10
16	06	-05	30	-05	-22	-05	-11	-24	-13	-17	06	-08	01	-13	81	—	00	14
17	06	-05	-04	03	04	05	-05	08	03	01	-06	04	-00	02	-03	00	—	01
18	04	06	08	-08	-20	-02	-09	07	04	-08	00	-01	15	02	10	14	01	—

TABLE 9. Correlation Table of union of M1 and M2 (the two digits after decimal point). Top 6 are the common features, followed by features from M1 only, followed by M2 only.

ID	Feature
1	Flock was thinned
2	Age at slaughter
3	Separate farm clothing for each house
4	Plastic overshoes for each house
5	Feeding equipment failures
6	Number of farm workers (cubic polynomial)
7	Coccidiosis
8	Specific breed
9	Side ventilation
10	Number of rodent baits
11	Pigs on the farm
12	Number of farm workers (quadratic polynomial)
13	Wall of concrete
14	Number of farm workers
15	Who cleaned the houses before last crop
16	Who disinfected houses before last crop
17	Broiler farm nearby
18	Water supply: borehole

TABLE 10. Features in Table 9.

between features in M1 and in M2. The importance of these experiments is that they show how cross-validation should be used to determine predictive power of the selected set of features and the confidence in the parameters of the model. These considerations, rather than goodness of fit should be the driving force in model design.

## 5. Conclusion

In this work we have shown how machine learning methods can be used to validate epidemiological models (via cross-validation) and to aid in their construction by suggesting features of interest (via feature selection methods). We plan to continue exploring potential applications of machine learning methods in epidemiology with more sophisticated methods and new real-life data.

## References

- [1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.
- [3] I. Dohoo, C. Ducrot, C. Fourichon, A. Donald, and D. Hurnik. An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. *Preventive Veterinary Medicine*, 29:221–239, 1996.
- [4] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Number Monograph 38 in CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1982.
- [5] N. M. Ferguson, C. A. Donnelly, M. E. J. Woolhouse, and R. M. Anderson. The epidemiology of BSE in cattle herds in Great Britain. II. Model construction and analysis of transmission dynamics. *Philosophical Transactions of the Royal Society of London*, 352(1355):803–838, July 1997.

- [6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
- [7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, New York, 2001.
- [8] P. Hermans, D. Fradkin, I. Muchnik, and K. Morgan. Prevalence of wet litter and associated risk factors in broiler flocks in the UK. Submitted to *Veterinary Record*, 2004.
- [9] C. Heuer, H. Luinge, E. Lutz, Y. Schukken, J. van der Maas, H. Wilmink, and J. Noordhuizen. Determination of acetone in cow milk by Fourier transform infrared spectroscopy for the detection of subclinical ketosis. *Journal of Dairy Science*, 84:575–582, 2001.
- [10] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1143, 1995.
- [11] B. Rockhill, D. Spiegelman, C. Byrne, and D. J. H. G. A. Golditz. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *Journal of the National Cancer Institute*, 93(5):358–366, March 2001.
- [12] N. Rose, J. Mariani, P. Drouin, J. Toux, V. Rose, and O. Colin. A decision-support system for salmonella in broiler-chicken flocks. *Preventive Veterinary Medicine*, 59:27–42, 2003.
- [13] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 2nd edition, 1998.
- [14] A. Wade. Derivation versus validation. *Archives of Disease in Childhood*, 83:459–460, 200.

DEPARTMENT OF COMPUTER SCIENCE, RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY, PISCATAWAY, NEW JERSEY 08854, USA

*E-mail address:* `dfradkin@cs.rutgers.edu`

DIMACS, RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY, PISCATAWAY, NEW JERSEY 08854, USA

*E-mail address:* `muchnik@dimacs.rutgers.edu`

UNIVERSITY OF LIVERPOOL, FACULTY OF VETERINARY SCIENCE, LEAHURST VETERINARY FIELD STATION, NESTON, CH64 7TE, UNITED KINGDOM

*E-mail address:* `hermans@liverpool.ac.uk`

UNIVERSITY OF LIVERPOOL, FACULTY OF VETERINARY SCIENCE, LEAHURST VETERINARY FIELD STATION, NESTON, CH64 7TE, UNITED KINGDOM

*E-mail address:* `K.L.Morgan@liverpool.ac.uk`