Public Health Risk-Based Inspection System

for

Processing and Slaughter

Appendix E – Data Analyses

APPENDIX E – DATA ANALYSES

The main text of this report outlines the method and algorithm Food Safety and Inspection Service (FSIS) is currently considering for a public health risk-based inspection system. When developing an algorithm to allocate FSIS resources based on public health risk, it is important to determine how the establishment's finished products, and the species and processes used in the establishment, could affect risk. That includes both the potential magnitude and probability of an establishment affecting public health. The data available on which the algorithm could be based are discussed in Appendix D. In this appendix, those data are examined and analyzed for use in assessing an establishment's public-health risk.

First, an analysis of the relative risks of the bacterial species/processes in the FSIS-requested expert elicitations is presented. This analysis is followed by an examination of production volume data. Noncompliance reports (NRs), food safety consumer complaints, food safety recalls, enforcement actions, *Salmonella* verification categories, ready-to-eat (RTE) *Listeria monocytogenes* Alternatives, and zero-tolerance pathogen test results are then examined. Each of those parameters was assessed for correlations and relationships to the other parameters that are considered indicators of a loss of process control and, therefore, a risk to public health. These analyses were conducted to examine both how well the individual parameters predict food safety contamination events (i.e., positive pathogen results), and how they are related to each other. The latter analysis can provide information on the interdependence and potential weighting of factors, if that was to have been done in the algorithm. Other establishment characteristics (age, square footage, number of employees, Hazard Analysis Critical Control Point [HACCP] training, use of chemical sanitizers, and the number of inspectors) are also evaluated.

RELATIVE RISK OF SPECIES/PROCESS

In order to rank the potential hazards of the products regulated by FSIS, the Agency has elicited the opinion of experts. Such "expert elicitations" have been conducted three times—in 2001, 2005, and 2007. The 2005 and 2007 elicitations were conducted in a similar manner, and are relevant to previous and current risk-based inspection proposals (RBI).

In this section, the consistency of the elicitation results across the various experts is assessed, both within a given elicitation and across the different elicitations, for scientific interpretation and application. It is also important to compare the results of the elicitation with the Agency's own microbial data, and to interpret the results in the context of published literature on food safety hazards. Summaries of those analyses and comparisons for the 2005 and 2007 elicitations are presented in this section. The relations between the elicitations and outbreak data are discussed in Appendix A.

Consistency of Expert Elicitations

Although there were differences in the worksheets and procedures used for the two recent expert elicitations, they are comparable enough to allow comparisons. Specifically, both expert elicitations included rankings of the relative risks of foodborne illness resulting from consumption of approximately 25 processed meat and poultry products. However, the 2007 elicitation included an additional product (thermally processed, commercially sterile meat and poultry), additional worksheets for ranking relative risks for vulnerable consumers and attribution of illness by pathogen to specific food types, and limited the rankings from 1 to 10 rather than allowing open-ended ranking. Analyses have been conducted to compare the 2005 and 2007 elicitations using the rankings for the 24 processed meat and poultry products common to both elicitations. The two elicitations were well correlated, with a Spearman correlation coefficient, "ρ," of 0.95. The strong positive correlation between the two elicitations of different experts provides confidence in the results of each expert elicitation.

Correlations between Expert Elicitation Results and Microbiological Data

The FSIS microbial sampling results can be analyzed to evaluate if those products and processes that were ranked in the expert elicitations as having the highest likelihood of illness are those most likely to have a contamination event. The control measures that are in place by industry might affect the actual incidence of contamination, but some confirmation of the rankings in light of actual FSIS data are possible. Therefore, the incidence of *Escherichia coli* O157:H7 (*E. coli* O157:H7), *Salmonella*, and *L. monocytogenes* in various end products has been compared with the expert elicitation risks for which we have data. Limitations in these analyses include matching the end products in the elicitations with product descriptions in the FSIS laboratory database, the low number of positive results for *E. coli* O157:H7 and *Lm* in the high-ranking products, and the fact that only a few of the ranked risks have consistent quality historical data available for analysis. Results for analyses conducted to date are included later in this appendix.

PRODUCTION VOLUMES

One component of the potential public health impact of a contamination event at an establishment is the production volume. One question that was raised by stakeholders was how accurately FSIS estimates of an establishment's production volume are. The FSIS has production volume data from a few sources: inspectors have provided information on the volumes of each product that FSIS-regulated establishments produce; for certain RTE products, industry provides volume data through an Office of Management and Budget (OMB)-approved survey; production volume from a random sample of FSIS-regulated establishments; and FSIS inspectors report production volume for ground beef when *E. coli* O157:H7 samples are collected.

The FSIS inspection force has, through Performance Based Inspection System (PBIS) extension data, provided production volume estimates for FSIS-regulated facilities. Details of how the inspectors estimate and record the volume in PBIS are presented in Appendix D. In order to assess how well the inspection force can estimate the volume, the inspector-generated results can be compared to other available data on production volume. Although industry data are not currently available for all establishments, industry-generated data for two subsets of FSIS-regulated establishments are available for analysis as follows: establishments subject to sampling under *L. monocytogenes* Alternatives participated in a mandatory OMB-approved information-collection program using FSIS Official Form 10,240-1, which includes a question on annual production volumes of different types of products; and a one-time OMB-approved voluntary survey that was conducted in order to obtain data needed for regulatory impact analyses, including production volume, from a random sample of FSIS-regulated establishments. These are compared below.

As part of the mandatory OMB-approved information collection related to *L. monocytogenes* Alternatives, industry provided volume data for a subset of establishments. The production volume figures collected under this program are called "10,240-1 volume data." This program requires annual OMB approval for continuous information collection. Since 2004, FSIS has requested establishments that produce post-lethality exposed RTE product to provide FSIS with estimates of annual production volume and related information for the types of RTE meat and poultry products processed. To facilitate compliance with this requirement, and to ensure that the information is collected in an efficient and uniform manner, FSIS has made available FSIS Form 10,240-1. A unique property of the 10,240-1 volume data is that the volume estimates are provided by industry as opposed to being estimated by FSIS inspectors for the same facilities. The purpose of this section is to compare the 10,240-1 production volume data provided by industry with those made by FSIS inspectors.

The program to gather FSIS inspector-generated volume estimates began in 2006, while 10,240-1 production volume data collection began in 2004. For the present study, the 10,240-1 volume data and the inspector-generated volume data will be compared for the year 2006. In filling out Form 10,240-1, an establishment only needs to update a previous year's production volume estimate if there has been a significant change in production volume. Thus, the 10,240-1 volume estimates for 2006 may contain estimates that were entered in 2004 or 2005, but have not been updated since the volumes produced by the facility have not changed significantly. Thus, some of the volume data in the 10,240-1 volume dataset may be labeled as 2004 or 2005 data, but actually represent 2006 data, since these entries are for volumes that have not changed.

Differences in the 10,240-1 and Inspector-Generated Volume Datasets

A major difference between the 10,240-1 and inspector-generated volume datasets is that the 10,240-1 data include only establishments that produce RTE products, while the inspector-generated data are for all FSIS-inspected establishments. However, the two datasets have in common establishments that produce RTE products.

Another difference is the categories of RTE food items reported in the two datasets. The 10,240-1 data have nine RTE categories, including such items as deli sliced, deli not sliced, hot dogs, fully cooked, and fermented. The inspector-generated data have four RTE categories, including RTE fully cooked 100 percent meat, other RTE fully cooked meat, RTE not fully cooked meat, and RTE 100 percent poultry. The only food category the two surveys have in common is the fully cooked category. However, the 10,240-1's fully cooked category includes only post-lethality exposed food items, while the inspector-generated data's fully cooked category includes fully cooked items that are both post-lethality exposed and those that are not post-lethality exposed. Thus, for the fully cooked category, the inspector-generated volume estimates should be larger than the 10,240-1 volume estimates.

There are several differences in how production volumes are reported in the 10,240-1 and inspector-generated volume datasets. The 10,240-1 volume figures are for a yearly volume, while the inspector's volume estimates are reported as falling in one of seven average daily volume ranges and five ranges for the average number of days per month the product is shipped. The product of these two variables places the average monthly product volume into one of 35 ranges of pounds of product produced/shipped in a month. In summary, associated with each facility in the 10,240-1 dataset is a single volume estimate representing the annual production volume at that facility. Associated with each facility in the FSIS dataset is a single volume range that brackets the monthly production volume at that facility.

Despite these differences, some comparisons between the 10,240-1 RTE volume dataset and the FSIS RTE volume dataset were made.

Comparison of 10,240-1 and Inspector-generated Volume Data

The 10,240-1 fully cooked RTE volume data (RTE fully cooked 100 percent meat plus other RTE fully cooked meat) were compared with the 2006 inspector-generated fully cooked RTE volume data. As mentioned above, the 10,240-1 fully cooked volume data represent yearly production volume, while FSIS fully cooked volume estimates are reported as falling in one of six daily volume ranges and five ranges for number of days per month the product is shipped. To facilitate comparison of the two datasets, the inspector-generated data was first converted to average monthly production volume by multiplying the midpoint of an establishment's average daily volume range by the midpoint of its range for average number of days per month the product is shipped. This average monthly production volume is then multiplied by 12 to obtain an estimate of the average annual volume produced.

A linear regression of the two datasets for the fully cooked 100 percent meat category (the only RTE food category the two datasets have in common) is presented in **Figure E-1**. The two datasets have 1,097 RTE establishments in common. The correlation coefficient (R) is 0.58. Notice that the 10,240-1 volume data are on average 0.492 times the inspector-generated volume data in the regression. This means that the inspector-generated volumes are about twice (1.0/0.492) as large as the volume figures collected through the Form 10,240-1. This difference can be partially explained by the fact that the inspector-generated volume estimates include both post-lethality exposed products and those that are not post-lethality exposed, while the 10,240-1 data only includes post-lethality exposed food items. However, the difference appears too large to be fully explained by this factor.

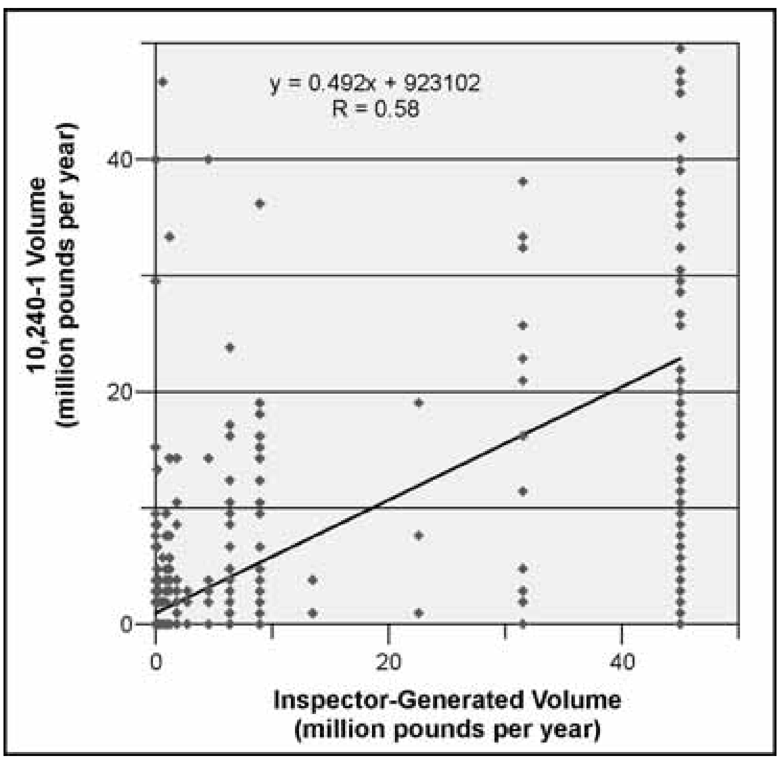


Figure E-1. Correlation Between 10,240-1 2006 and Inspector-Generated 2006 Volume Data for Fully Cooked Products.

In the above analysis, the inspector-generated volume data are the midpoints of 35 ranges. Thus, there are only 35 values that these volume data can assume. The original 10,240-1 volume data can be any number and are thus not constrained by this restriction. To examine if this constraint difference is the source of the low correlation in Figure E-1, we transformed 10,240-1 data to have the same constraint as the inspector-generated data. Each 10,240-1 volume datum was mapped into the appropriate range of the 35 volume categories, and assigned the midpoint of that range**. Figure E-2** presents the correlation of these two datasets after the transformation.

As can be seen above, the correlation is not greatly improved. The new correlation coefficient is R = +0.6089.

The 10,240-1 volume data provided by industry and the volume data estimated by FSIS inspectors have a fairly good positive correlation. However, there is also a high degree of variation between the two datasets. The coefficient of determination is $R^2 = 0.3707$, which shows that the inspector-generated volume data account for about 37 percent of the variation found in the 10,240-1 volume dataset.

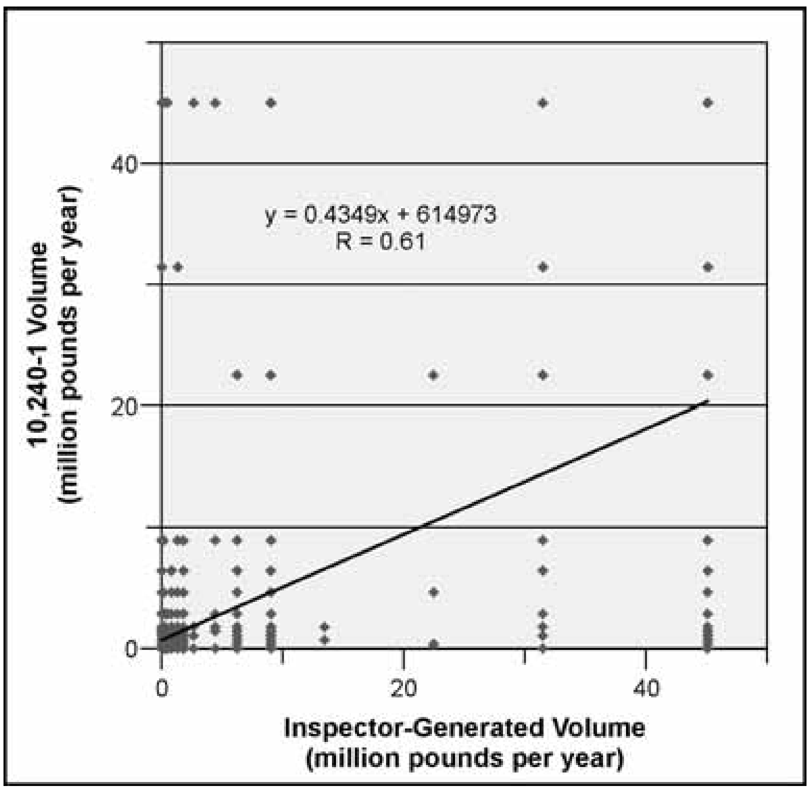


Figure E-2. Correlation Between the Transformed 10,240-1 Volume Data and Inspector-Generated Volume Data for Fully Cooked Products During 2006.

Comparisons Among Years for 10,240-1 RTE Volume Data

In this section and the following section, the consistency of the 10,240-1 RTE volume datasets is evaluated by comparing them among years 2004 to 2007. The 10,240-1 2006 database was created in late December 2006. In early 2007, FSIS asked industry to provide new estimates of production volume. In this data call, every RTE establishment was asked to enter a volume estimate regardless of whether its production volumes had changed or not. Thus, every 2007 entry in the 10,240-1 volume dataset was entered in early 2007. Since the 10,240-1 2006 volume survey was up-to-date as of the end of December 2006 and the 10,240-1 2007 volume survey data is from early 2007, one might expect that there would be little change in the two industry-provided estimates of RTE production volume.

The 2006 10,240-1 volume dataset has data on 4,930 RTE production establishments, while the 2007 10,240-1 volume dataset has data on 1,677 (data in the 2007 10,240-1 survey represent RTE establishments that had responded to the FSIS data call by July 2007). The two datasets have 976 RTE production establishments in common. **Figure E-3** presents a correlation between the two datasets with one outlier removed. The correlation coefficient is R = 0.65. If the one outlier is included, the correlation coefficient between the 10,240-1 2006 and 10,240-1 2007 volume estimates is R = 0.071.

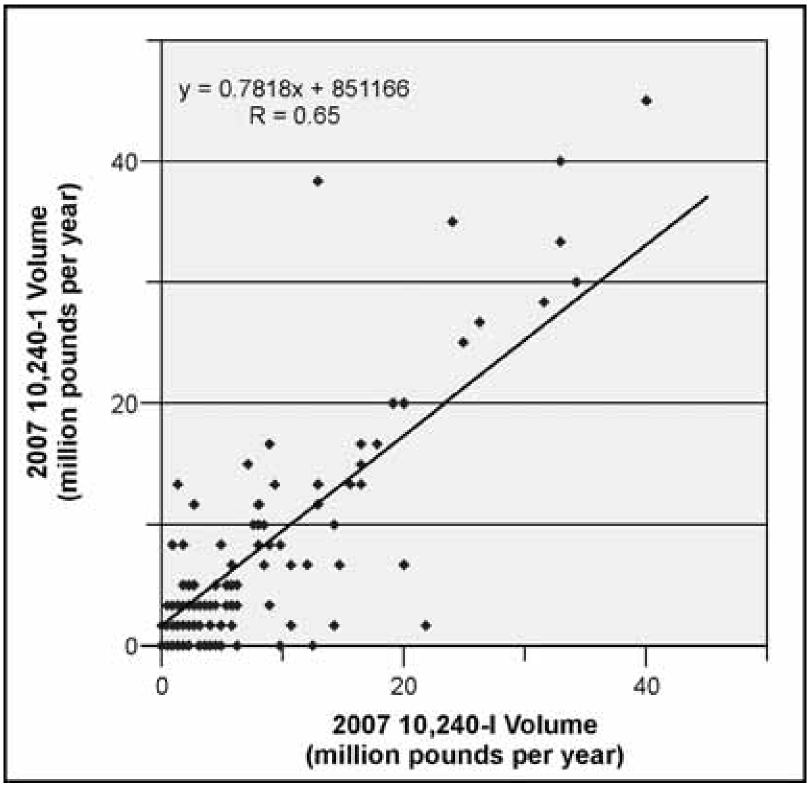


Figure E-3. Correlation Between 10,240-1 2006 and 10,240-1 (2007 Volume Data)

As can be seen from the Figure E-3, the 10,240-1 2007 RTE production volume estimates are larger than the 10,240-1 2006 volume estimates by a factor of about 1.3.

The average absolute difference in volume estimates between 10,240-1 2006 and 10,240-1 2007 is 1.7 million pounds of fully cooked RTE product per year per establishment.

Updating of 10,240-l Volume Data

The 10,240-1 volume estimates for 2006 contain RTE production volume estimates that were entered in 2004 or 2005, but have not been updated since the volumes produced by the facility have not changed significantly. **Table E-1** presents the number of RTE establishments with 2004, 2005, and 2006 volume estimates.

Table E-1. Number of Establishments with Given Entry Year in 10,240-1 2006 Volume Dataset

Year	Number of Establishments	Percent
2004	1,503	61.78
2005	754	30.99
2006	174	7.55

In total, there are 2,439 establishments in the 10,240-1 2006 database. Six establishments in the database did not have a date of entry. Table E-1 demonstrates that 62 percent of the establishments have not updated their volume estimates since 2004, and 31 percent have not updated their volume estimates since 2005. Only 8 percent of the establishments entered new volume estimates in 2006. Presumably, this means that the majority of establishments have not changed their production volume in the past 2 years.

The FSIS is looking for potential methods or additional means to compare the 10,240-1 and inspector-generated volume data, including having Enforcement, Investigation, and Analysis Officers (EIAOs) report more detailed information on product- and processing-specific volumes when they conduct food safety audits. Having the EIAOs gather that information would not only facilitate the comparison between the volume data provided by industry with that captured by FSIS field personnel, but would also provide means for independent verification of the volume data captured by the FSIS inspection force for a random sample of establishments.

Comparison of Voluntary Industry Survey and FSIS Data

The second OMB-approved survey mentioned above is a voluntary survey of FSIS-regulated establishments; in that survey, industry supplied data on production volume (Cates et al. 2006). The purpose of the voluntary survey was to collect uniform information on practices and technologies used to control pathogens and promote food safety in the meat and poultry industries. In addition to collecting information on practices and technologies, the survey collected information on establishment characteristics including the volumes and types of products produced. The survey sample was stratified by inspection status (Federal versus state) and HACCP size (large establishments with 500 or more employees, small establishments with 10 or more but fewer than 500 employees, and very small establishments with fewer than 10 employees and less than $2.5 million in annual sales). For Federally-inspected establishments, the universe includes 4,266 establishments from which a starting sample of 1,086 establishments was drawn. The sample design specified the sample size to yield precision of ±5 percent or better for estimates of all proportions, assumed a 90 percent eligibility rate for very small and small Federally-inspected establishments and a 95percent eligibility rate for large establishments, and assumed a target response rate of 75 percent.

The survey respondents provided production volume information by selecting a range of annual volumes (e.g., 10,000 to 49,999 pounds per year) for each type of meat or poultry product (beef, pork, other meat, chicken, turkey, and other poultry). The respondents also indicated the percentage of each type of meat or poultry product across eight product types (e.g., raw, ground and raw, not ground). The responses from these sets of questions were used to calculate ranges of production volumes for each meat and poultry product type for each establishment.

The industry-supplied data from the voluntary survey was then compared to inspector-generated volume data to assess how closely inspector-generated volume data matches industry-supplied volume data. The FSIS contracted with RTI International to conduct correlation analyses comparing the industry-supplied volume data to inspector-generated volume data.

To conduct the analysis, the product categories from the inspector-generated data were matched to the product categories in the voluntary establishment survey. Separate comparisons were made by individual product category (17 categories in total). In both datasets, volume data were collected as ranges of pounds produced (e.g., 10,000 to 49,999 pounds) over a specified time period. However, the ranges of pounds used for the responses differed between the two data sources, and the timing of data collection differed. For FSIS inspector-generated data, the time period referred to a one-month period during the first half of 2007; for the industry-supplied volume data, the time period referred to the amount produced in the "past year" relative to when the survey was administered over the July through November 2005 period. Because of the differences in the response ranges used for the volumes in each data source, the comparisons were made by determining whether the ranges of volumes from each of the data sources overlap. Prior to making the comparisons, data from each source were transformed as described below.

First, for the FSIS inspector-generated volumes for each establishment and product category, a range for the annual number of days of production was computed by multiplying the minimum and maximum number of days the product was produced over the prior 30 days by 12. Then, the minimum annual days was multiplied by the minimum daily production volume to get a minimum annual production volume, and the maximum annual days was multiplied by the maximum daily production volume to get a maximum annual production volume. This provides an absolute annual range by product category.

For the voluntary survey volumes, the percentage of production by product category (e.g., raw, ground; raw, not ground; thermally processed, commercially sterile) was multiplied by the minimum and maximum total annual production volumes to obtain a minimum and maximum annual volume for each product category-species combination.

Establishments in the two datasets were then matched using the FSIS establishment numbers for each product category. The voluntary establishment survey included volume data for relevant processed meat and poultry products for 570 establishments, most of which produced multiple products. For each comparison, it was first determined whether both datasets reported a volume for each product category, and then whether the volume ranges from each of the datasets overlapped.

The results of the analysis are shown in **Table E-2**. The ranges from the self-reported volumes from the voluntary establishment survey overlapped with the ranges from the FSIS inspector-generated data about two-thirds of the time. However, in many cases, establishments reported volumes on the voluntary survey for products for which the FSIS inspector data did not indicate a volume. This is likely because of the seasonality of production of certain products—that is, some products that an establishment produces over the course of a year were not produced during the month of the FSIS inspector survey. Other reasons for differences in whether both datasets included a volume for a particular product category and whether the ranges overlapped could be due to the difference in the time period of the surveys as described above (approximately 2 year's difference) or that the definitions of the product categories were slightly different in each dataset.

Table E-2. Comparison of Processed Meat and Poultry Volumes Generated by FSIS Inspectors in 2007 and Volumes Collected on a Voluntary Industry Survey in 2005 (570 establishments)

Product Category	No. Establishments with FSIS Inspector Volume	No. Establishments with Voluntary Survey Volume	No. Establishments with Volumes in Both Datasets	No. Establishments with Overlapping Ranges	Percent of Establishments with Overlapping Ranges
Raw Intact Beef and Raw Beef Trimmings	169	180	148	84	57%
Raw Intact Pork	156	166	118	81	69%
Raw Intact Other Meat	40	63	0	---	---
Raw Ground Beef	127	171	119	76	64%
Raw Ground Pork	125	174	107	72	67%
Other Raw Ground Meat	20	37	6	3	50%
Fully Cooked Meat	250	298	219	158	72%
RTE Not Fully Cooked Meat and Poultry	58	48	15	10	67%
Raw Intact Chicken	101	117	76	43	57%
Raw Intact Turkey	18	34	12	9	75%
Other Raw Intact Poultry	3	9	1	1	100%
Raw Ground Chicken	18	45	12	6	50%
Raw Ground Turkey	7	27	6	2	33%
Other Raw Ground Poultry	2	2	0	---	---
RTE Poultry	120	207	108	63	58%
Partially Cooked Meat and Poultry	92	124	70	46	66%
Thermally Processed Commercially Sterile Meat and Poultry	16	23	13	11	85%
Total	**1,322**	**1,725**	**1,030**	**665**	**65%**

Based on the results of this analysis, the voluntary survey data provide a moderate degree of validation of the inspector-generated volumes. However, the match rates would likely have been higher if the time period were the same, the lengths of time included in the volume estimates were the same, and the product definitions were defined exactly the same. This analysis does provide some confidence in the PBIS data, especially given the proposed categorization of the volume data for use in ranking public-health risk, as discussed in the main text of the report.

In addition to the questions about the ability of the FSIS inspection force to collect accurate information on production volume, some stakeholders have questioned whether production volume should be a component of an establishment's inherent risk regardless of its accuracy. The argument used is that there might not be any correlation between production volume and a lack of process control that could put the public's health at risk, or that large-volume establishments might have even better control measures in place and, therefore, pose less risk to public health.

It is important to note, however, that even if large-volume establishments are no more likely or even less likely to have lost control of its food safety system, establishments that produce larger volumes of product have a greater potential to impact public health—that is, the more servings an establishment produces, the more people who could potentially consume the product. Therefore, FSIS uses production volume as a surrogate or measure of consumption of an establishment's product and, therefore, an indicator of potential magnitude exposure. Therefore, as a matter of policy, FSIS believes that volume must play a role in risk-based inspection, and the lack of a correlation between volume and loss of process control (or the presence of an inverse correlation) should not dictate whether volume is taken into account in an public-health risk-based algorithm.

Despite that caveat, FSIS does believe that examining the relationship between establishment production volume and indicators of establishment performance is valid, not only to address stakeholders' questions, but also to assist the Agency in focusing outreach activities in addition to inspection resources (e.g., if establishments with a given production volume have poorer performance, FSIS could focus its outreach activities to establishments in that category). With those purposes in mind, FSIS conducted analyses comparing production volume with microbial sampling results, and other indicators of an establishment's food safety performance that have been proposed previously for use in risk-based inspection (NRs, consumer complaints, recalls, and enforcement actions). The results of those analyses are presented later in this appendix.

Public Health NR Rates

Public-health-related NRs are a component of the currently proposed method for allocating resources as an indication of an establishment's control of its food safety system, and subsequent potential public health significance. The NRs are discussed in more detail in Appendix D. In this section, the categorization of those NRs according to potential relation to public health is further examined by looking at the correlations between NRs and other potential indications of process control such as pathogen results, consumer complaints, recalls, enforcement actions, and *L. monocytogenes* Alternative. These analyses provide insight as to whether NRs, or subsets of NRs, are indicators of an establishment being more likely to have a loss of food safety control and, therefore, their importance as a component of public health risk-based inspection.

NRs and Pathogen Test Results

In order to determine if the expert opinion used to identify the most important public-health-related NRs is valid, analyses have been conducted to see if a specific subset of NRs are more predictive of an establishment's performance than others. The analysis evaluated several subsets of NRs (e.g., facility NRs, sanitation NRs, or HACCP NRs) to determine which were better predictors of Salmonella, E. coli O157:H7, or *L. monocytogenes* test results. These analyses were conducted by product types (i.e., data are used only for the products that are tested for a given pathogen).

One issue that was raised by stakeholders in previous analyses was that some NRs are based on an inspector's opinion and not a quantitative measure. Another issue raised was that not all NRs are directly related to process cleanliness. These analyses have been conducted using several different subsets of NRs in order to address these two issues. By looking for statistical correlation with known events, FSIS can determine which NRs are the best indicators of the loss of process control.

NRs are defined as violations of regulations as recorded in the PBIS. The FSIS inspectors have recorded violation information on establishments in PBIS for several years. Test results for pathogens in meat and poultry products are similarly recorded in a system called M2K. The question to be asked of the data then is, “Can we reliably predict future M2K positives (presence of pathogens in an establishment) based on the observation of recent establishment performance (as measured by PBIS NRs)?”

To answer this question effectively, lift statistic is adopted. Here “Lift” is defined as the ratio of “the number of cases of M2K positives after PBIS NRs” to “the total number of cases of M2K positives regardless of PBIS NRs.” The concept of lift statistic is explained in more detail later in this appendix.

Lift is a measure that indicates how much more likely it is, on average, for an establishment to have positive pathogen test results if it has also failed inspection(s), versus having such issues without taking into account inspection results. By computing the lift for various subsets of NRs, subsets of establishments, timeframes, and pathogens, FSIS can find any combinations that produce a strong predictor of pathogen presence and, therefore, could be candidates for incorporation into the RBI algorithm.

The M2K and NR are daily data, and it is desirable to examine their correlations not only among the same day occurrences but also occurrence aggregations over consecutive multiple days, which is called “time window.” The framework of time windows, as described in Figure 5-13, allows flexibility in answering various types of questions. In the case of relationship of NR versus *Salmonella* in M2K, the aggregation time window of NRs proceeds that of *Salmonella* in M2K, since FSIS interested in knowing how NRs are predicative of *Salmonella* in M2K. The time window is a dynamic variable, in which domain changes as a viewpoint changes. Thus, for each viewpoint, the number of NRs and the number of pathogen positives are found in a particular time-window to be used to compute a lift. The “Overview of Analytic Methodology” section later in this appendix describes lift and how it is calculated.

Figure E-4 illustrates the results of analyses for three NR subsets against positive findings of Salmonella in M2K. In this case, all establishments were included. The y-axis shows the computed lift. The time window into which the PBIS violations were aggregated is shown on the x-axis. The aggregation timeframe is referred to as the “evidence window size.” If any NRs were found in that timeframe, then the analysis looked ahead for 14 days to determine if any tests reported positive for Salmonella. The three subsets of NRs analyzed were: all NRs, only NRs in the set proposed by the industry coalition, and only NRs of type 3 (previously identified as public-health-related NRs). The bars indicate 95 percent randomization confidence intervals for each point.

Lift values higher than 1.0 indicate a positive correlation between the occurrences of positive pathogen results and the observed violations. Lift values equal to 1.0 represent a null hypothesis of no correlation. From Figure E-4, observing at least one occurrence of Type 3 NR over the past 7 days increases by threefold, on average, the chance of recording a positive result of *Salmonella* test over the following 2 weeks (with respect to the baseline expectancy that does not take into account any violations). This result can be seen as a relatively strong indication of the potential utility of these violations in predicting adverse outcomes of microbial testing. In other words, given the evidence collected in historical data, empirically, the risk of failing a test for *Salmonella* is substantially elevated at establishments that recently were found to be noncompliant.

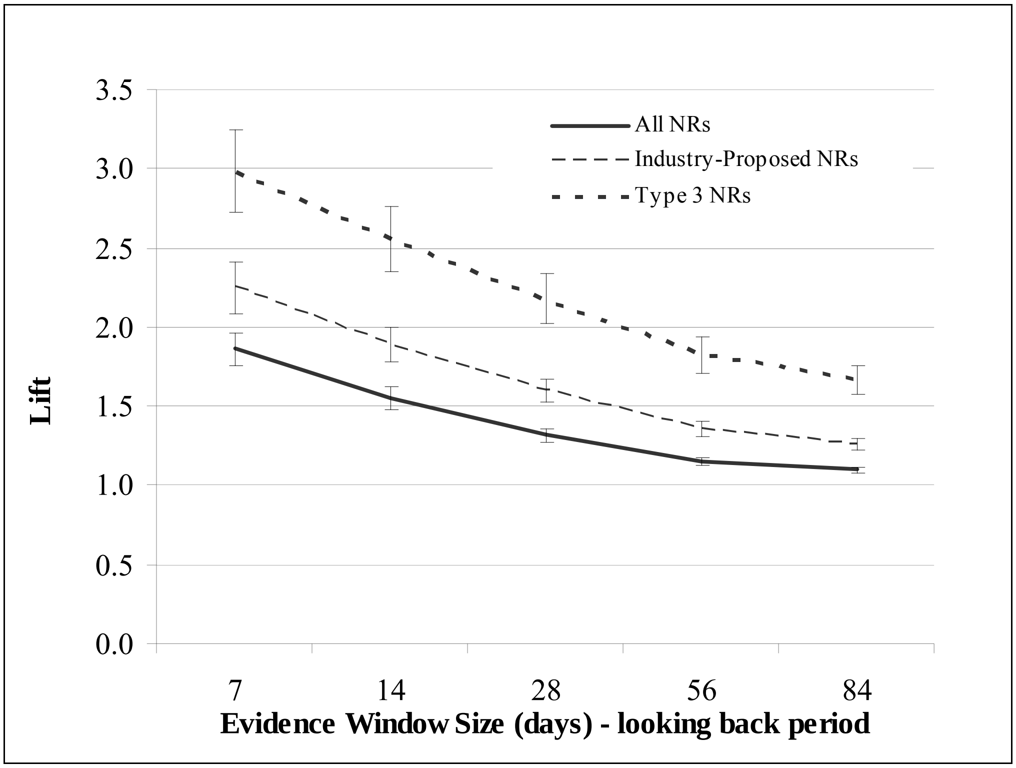


Figure E-4. Lift Analysis Results for NRs Versus *Salmonella*.

Figure E-4 shows that for all evidence window sizes considered, the industry coalition subset of NRs is a better predictor of positive results of *Salmonella* tests than simply using all NRs, and using only the public-health-related NRs (Type 3) produces even better results. The observed differences are significant as suggested by the nonoverlapping confidence intervals depicted in the graph. The graph also shows that as the time window for aggregation becomes longer, the predictive ability of each NR subset declines. This is logical because the long aggregation periods blur possible correlations between NRs and the presence of pathogens (over long periods almost all establishments experience some positive pathogen results). A hypothesis test was conducted for the Null Hypothesis, H_0: Lift = 1.0 (no correlation between NRs and *Salmonella* positives), with data randomized (1,000 datasets, including the one original dataset). The randomization method is explained later in this appendix. The results show that lift values are significantly greater than 1.0 at p-value of 0.001 for all the randomized data.

The data are also used to generate Receiver Operating Characteristic (ROC) curves. The ROC curves shown in **Figure E-5** have been obtained for the same NR subsets by varying one of the parameters of the lift method: the size of the evidence window, while keeping the outcome window size constant at 14 days. The vertical axis corresponds to the rate of true positive predictions (sensitivity) and the horizontal axis denotes the rate of false positive predictions (1.0 – specificity). ROC curves are often used to evaluate predictive accuracy of classifiers or event detectors and they provide a convenient way of optimizing parameters of the models given the costs of different types of errors (false positives and false negatives). Curves that bend most strongly toward the upper left of the graph are considered to represent better predictive models.

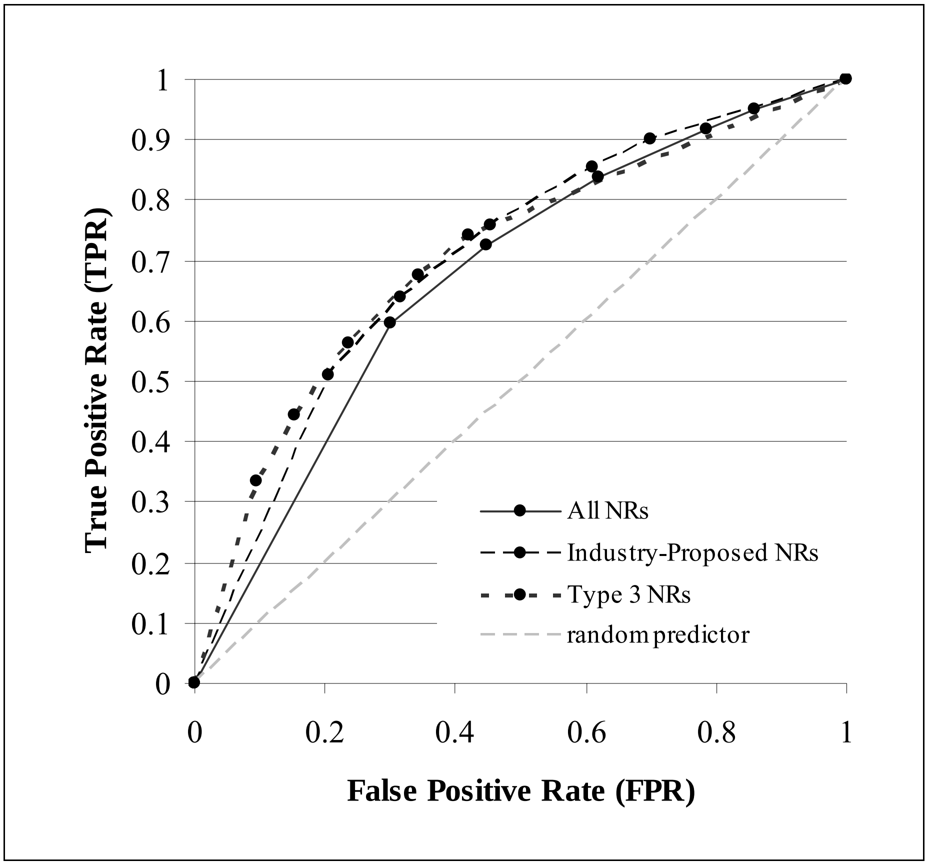


Figure E-5. ROC Curves for NRs Versus *Salmonella*

The area under an ROC curve (abbreviated AUC) is commonly used as measure of the overall capability of a model to discriminate classes of the output variable (i.e., either a positive or negative result of a test for *Salmonella* recorded within the outcome window). This is a more general evaluation of predictive utility than lift, since it directly takes into account a model's accuracy in predicting negative as well as positive outcomes. Lift focuses primarily on measuring utility in predicting positive outcomes. The simplest possible model would always predict the most frequent class of the output variable regardless of any available input variables. It would correspond to either the lower left or the upper right corner of the ROC diagram. In this example, this would be the former of the two denoting a model that always predicted a lack of positive pathogen results (without regard to NRs), since this is by far the most common occurrence within the data (i.e., on most days, most establishments are pathogen free). A model based on chance which picks predictions randomly according to the observed frequencies of test outcomes would result in a ROC curve identical with the diagonal connecting the lower left and upper right corner of the graph, and its AUC score would equate to 0.5. The perfect predictor would have AUC of 1.0, and in practice we expect a "fair" predictor to score at 0.7 or higher, although even a slight but significant departure from 0.5 does indicate some predictive power of the model and, therefore, some utility of the involved input variables. **Figure E-6** shows the AUC scores for each NR subset and the corresponding 95 percent randomization confidence intervals, obtained from the ROC curves shown in Figure E-5. Randomization tests identify all those values to be significantly greater than 0.5 at the p-value of 0.001.

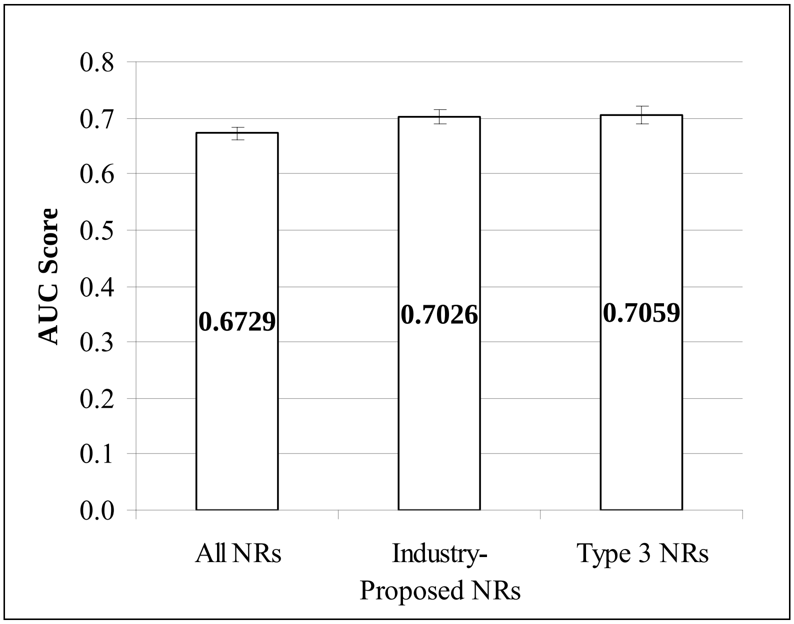


Figure E-6. AUC Scores for NR Subsets for *Salmonella*

A similar analysis was also performed for *E. coli* testing and positive events. *E. coli* positive results are much sparser than in the case of *Salmonella* records. This scarcity of positive results makes the analysis more difficult as can be seen in **Figure E-7**. Note that the lift values still tend to increase with higher specificity of the NR definitions and with shorter evidence window widths, but their estimates bear much less confidence than in the case of *Salmonella*. As with *Salmonella*, several tests were run to determine the optimum outcome window size based on the available historical data. In this case the optimum windows size was found to be 28 days. They are also less statistically deterministic, having p-values under the 0.05 threshold only for shorter evidence window widths.

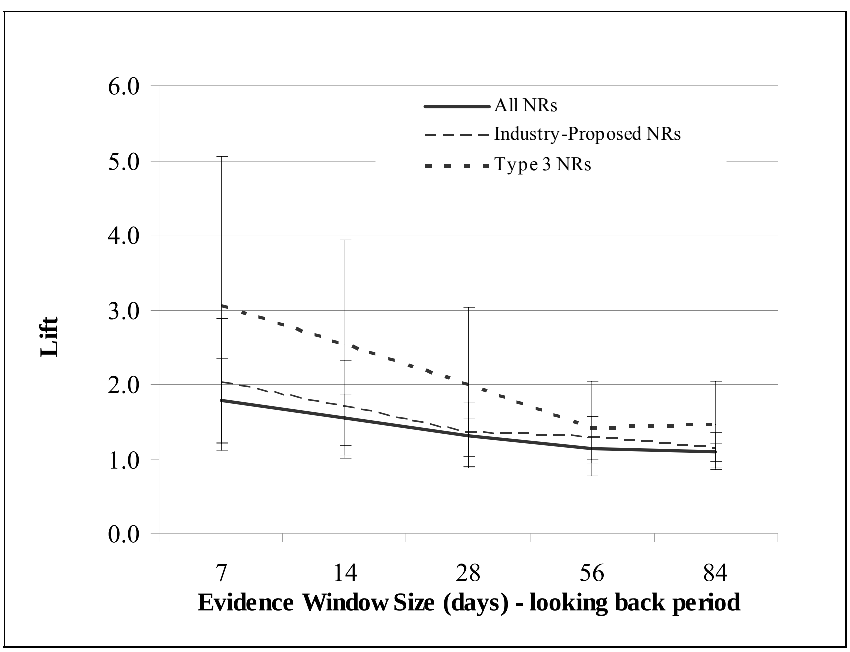


Figure E-7. Lift Analysis Result for NR Subsets Versus *E. coli* Positive Events; Outcome Window Size is 28 Days

The AUC scores obtained for *E. coli* data are also not as high as in the case of *Salmonella*. In this case, the most accurate predictor seems to be the subset using the least specific definition of NRs ("All"). However, the data are not strong enough to confidently consider it better than the other two results.

Two additional analyses were performed using the same methodology as above: one for *L. monocytogenes* and another with all pathogens (*Salmonella*, *E. coli*, and *L. monocytogenes*) combined under RTE projects. The RTE projects are presumably focusing on establishments that produce RTE products. The following codes are used in scoping out the pathogen tests and establishments falling under RTE projects ALLRTE, INTCONT, INTPROD, RTE001, and RTERISK1. Results for those two analyses are very close to each other. This maybe due to the fact that the establishments in *L. monocytogenes* pathogen tests and those under RTE projects are almost identical. Additionally, the majority of the positives of both analyses are from the same source—that is, *L. monocytogenes* pathogen tests under RTE projects (see later in appendix). Both sets of analysis yielded weak correlations. The observed lifts, as well as AUC scores were found to be statistically insignificant. **Figures E-8 (a)** and **(b)** show ROC curves for NRs versus *L. monocytogenes* positives, and all pathogen positives under RTE projects, respectively, for selected outcome window size. Similarly, **Figures E-9 (a)** and **(b)** show AUC score for those two analyses.

NRs and Food Safety Consumer Complaints

The issuance of NRs by FSIS inspection personnel are based upon an observed noncompliance during a scheduled inspection task and are associated with a certain regulatory citation. Consumers who experience problems with FSIS-regulated food products are able to register complaints and these complaints are monitored via a system known as the Consumer Complaint Monitoring System (CCMS). Not all complaints can be associated with a particular establishment. Some subset of NRs may be predictive of the occurrence of a particular subset of food safety consumer complaints. This analysis may aid in evaluating whether NRs that have been issued have any correlation to documented food safety consumer complaints that have been associated with individual establishments.

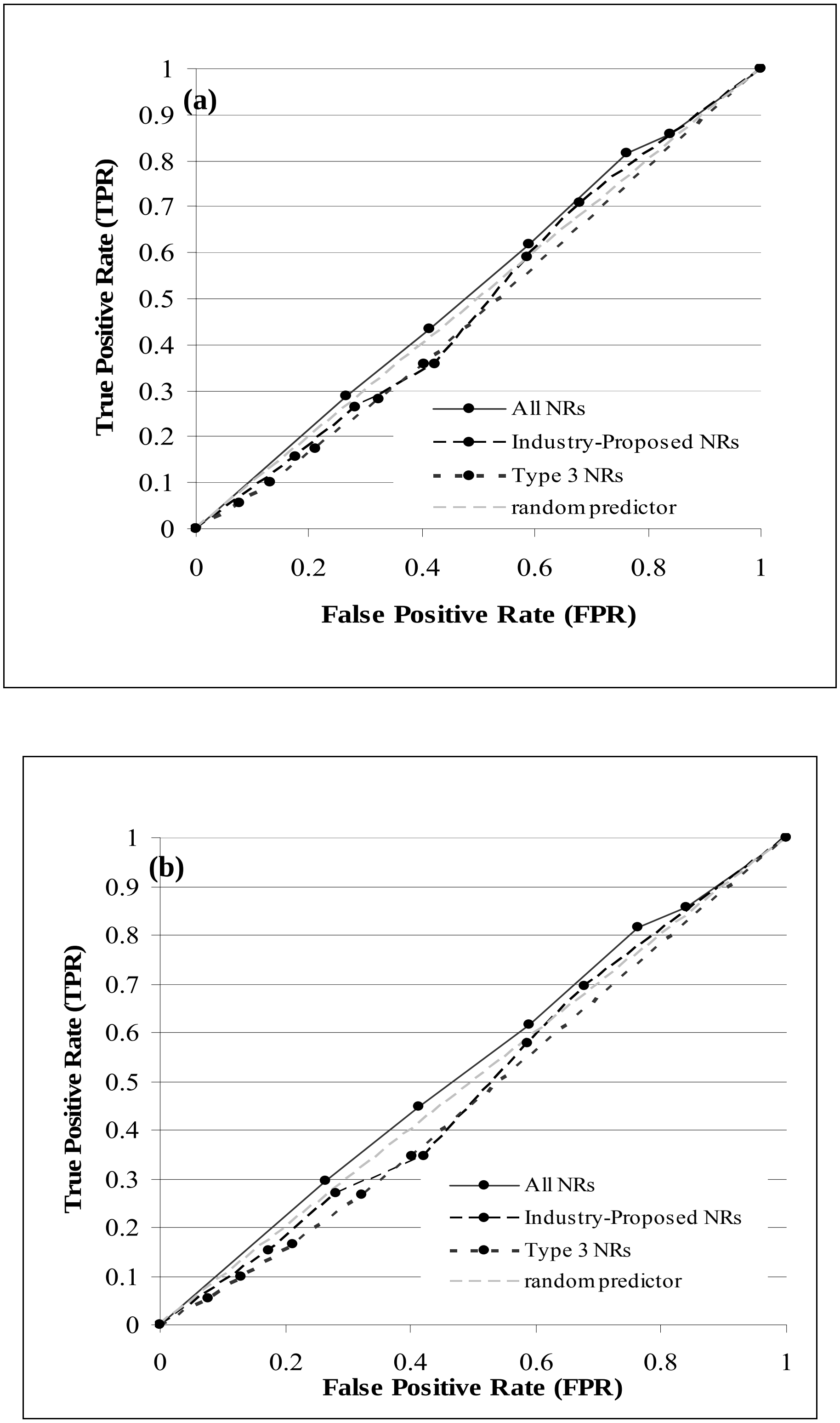
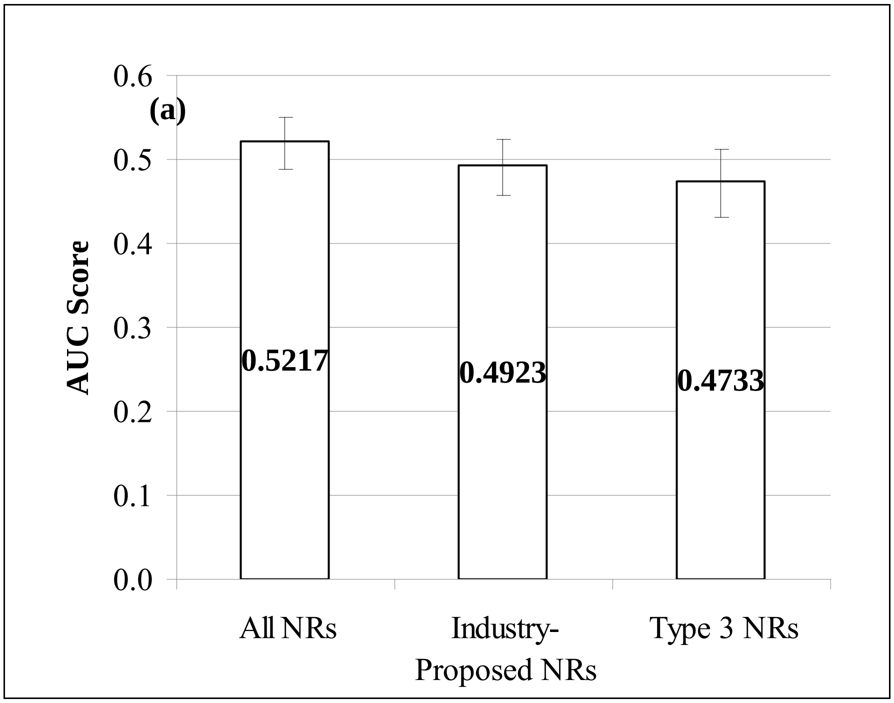


Figure E-8. ROC Curves for NRs Versus (a) *Listeria monocytogenes* Positives, and (b) All Pathogen Positives in RTE Products; Outcome Window Size is 7 Days.



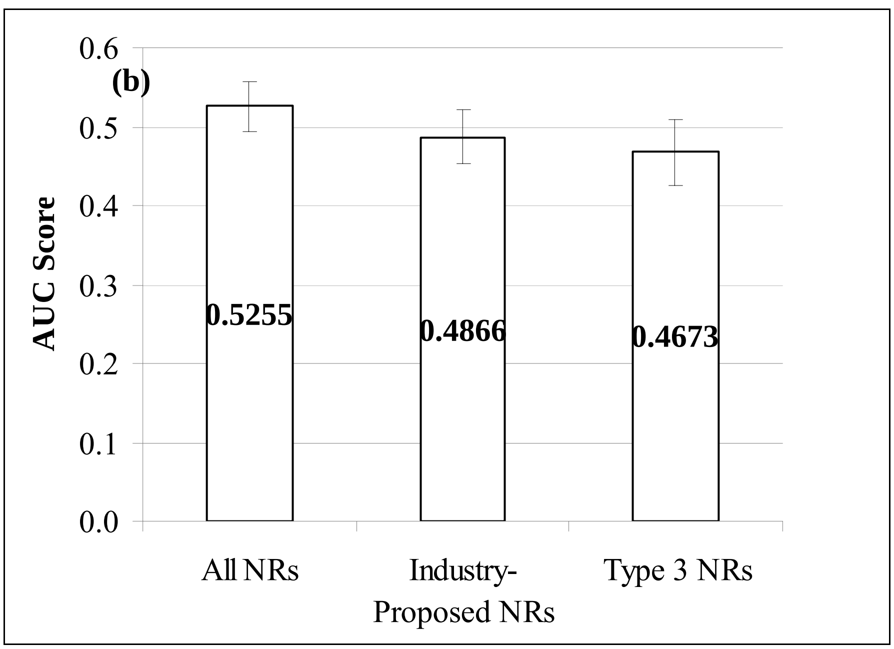


Figure E-9. AUC Scores for NRs Versus (a) *Listeria monocytogenes* Positives, and (b) All Pathogen Positives in RTE Products; Outcome Window Size is 7 Days

Analyses examining that relationship returned a few indications of possible correlation, but very few of these results can be considered statistically significant. A similar methodology was utilized in this work as employed above where lift was computed for various windows sizes and randomization performed to validate results. It was found that using PBIS Type 3 noncompliance records to predict a set of CCMS events provided by the USDA FSIS Office of Program Evaluation, Enforcement, and Review (OPEER) using an 84-day evidence window width (i.e., the time period over which the NRs were aggregated) and 28- and 56-day outcome window widths (the timeframe to look forward for complaints) yields lifts of 1.115 and 1.12, respectively. P-values obtained from significance tests for these lifts are 0.043 and 0.028. However, the lower limits of the 95 percent confidence intervals obtained through hypothesis test using bootstrap randomization for these values of lift are below 1.0. This may indicate low robustness of those results to random sampling of the establishments. Type 3 noncompliances are apparently also potentially useful in predicting CCMS epidemiological (EPI) events when using either 56- or 84-day evidence window widths and 28- or 56-day outcome window widths. These analyses yielded statistically significant lifts ranging from 1.38 to 1.5 (with the same caveat regarding lower confidence limits as above). The only significant results based on Industry Coalition definition of NRs correspond to CCMS OPEER cut events and outcome window width of 28 days, with evidence window widths of either 14 or 28 days. The resulting lifts stand at merely 1.08 (albeit statistically significantly greater than 1.0 and with the lower confidence limits also greater than 1.0). The predictive value of these NRs therefore appears to be marginal. Randomization tests were performed to determine the upper and lower limits of 95 percent confidence intervals (95 percent rCI). A complete explanation of this methodology is included later in this appendix. In every case 1,000 randomization tests were performed to determine confidence intervals. These results are summarized in **Table E-3**.

Table E-3. Relationship Between NRs and Food Safety Consumer Complaints

NR Type	Consumer Complaint	Windows, Days		Lift	95% rCI		p-value
		Evidence	Outcome		Lower	Upper	
Type 3	OPEER	7	28	0.9713	0.83097	1.10954	0.605
Type 3	OPEER	14	28	0.9632	0.83092	1.09198	0.68
Type 3	OPEER	28	28	0.9766	0.85437	1.09118	0.593
Type 3	OPEER	56	28	1.051	0.92667	1.18301	0.226
Type 3	OPEER	84	28	1.1153	0.96537	1.26109	**0.043**
Type 3	OPEER	7	56	1.0188	0.89126	1.13504	0.436
Type 3	OPEER	14	56	1.0204	0.90051	1.14153	0.414
Type 3	OPEER	28	56	1.0483	0.94217	1.15974	0.227
Type 3	OPEER	56	56	1.1062	0.98128	1.23181	0.052
Type 3	OPEER	84	56	1.1204	0.99025	1.25778	**0.028**
Type 3	EPI	7	28	0.7244	0.40552	1.092	0.796
Type 3	EPI	14	28	0.9417	0.5547	1.37042	0.577
Type 3	EPI	28	28	1.269	0.69829	1.88714	0.156
Type 3	EPI	56	28	1.4318	0.83836	2.0662	**0.043**
Type 3	EPI	84	28	1.4517	0.58705	2.19415	**0.031**
Type 3	EPI	7	56	1.0864	0.64408	1.53836	0.373
Type 3	EPI	14	56	1.1719	0.65518	1.6601	0.234
Type 3	EPI	28	56	1.2934	0.78637	1.85991	0.12
Type 3	EPI	56	56	1.3781	0.8196	1.93293	**0.038**
Type 3	EPI	84	56	1.5087	0.65818	2.28424	**0.016**
Industry-proposed	OPEER	7	28	1.0903	0.99264	1.1839	0.071
Industry-proposed	OPEER	14	28	1.0848	0.99344	1.17181	0.056
Industry-proposed	OPEER	28	28	1.0835	1.00061	1.17099	**0.033**
Industry-proposed	OPEER	56	28	1.0263	0.94552	1.1046	0.284
Industry-proposed	OPEER	84	28	1.035	0.96007	1.11284	0.179

NRs and Food Safety Recalls

A food safety recall may be triggered by a variety of factors once the product has entered commerce. The recall is classified based upon the relative health risk, and a Class I recall is a situation where the product has a *reasonable* probability of causing a health risk if eaten. Analyses of a subset of NRs, as they correlate to historical Class I recalls, may be predictive of an establishment's likelihood of experiencing a future recall.

Analyses examining that relationship highlighted two correlations as statistically significant. The first significant correlation involved predicting a Class I or Class II recall over ane outcome window 14-days-wide using the occurrence of any NRs over the period of the preceding 14 days. The second involved using the occurrence of Industry Coalition defined NRs over the previous 14 days to predict Class I or Class II recalls over outcome window sizes of 7 days. The computed lifts equal 1.28 and 1.42, respectively, and the p-values obtained from the randomization test of significance were 0.047 and 0.029. However, these results, summarized in **Table E-4**, do not appear robust against the random selection of establishments since the lower 95 percent confidence bounds do not exceed the value of lift=1.0.

Table E-4 Relationship Between NRs and Food Safety Recalls (Classes I and II)

NR Type	Windows, days		Lift	95% rCI		p-value
	Evidence	Outcome		Lower	Upper	
All NRs	7	14	1.3065	0.90616	1.76123	0.064
All NRs	14	14	1.2814	0.95699	1.61536	**0.047**
All NRs	28	14	1.1406	0.86667	1.41045	0.138
All NRs	56	14	1.0246	0.80316	1.24399	0.41
All NRs	84	14	1.0709	0.86706	1.25979	0.22
Industry-proposed	7	7	1.214	0.72991	1.80659	0.212
Industry-proposed	14	7	1.4234	0.95284	1.97039	**0.029**
Industry-proposed	28	7	1.2346	0.855	1.59726	0.108
Industry-proposed	56	7	1.0063	0.72345	1.30648	0.512
Industry-proposed	84	7	1.0878	0.84004	1.3283	0.274

NRs and Enforcement Actions

Enforcement actions are another indicator of an establishment's performance and may be considered to be a holistic indication of the efficacy of their process control system. Enforcement actions indicate serious or repeated violations and can include letters to the establishment, detention of product, or revocation of the inspection mark (effectively stopping all production). Analyses of a subset of NRs to determine if they correlate to enforcement actions and if they might be predictors of an establishment's food safety system design were conducted using a similar methodology as described in the preceding paragraphs. Only one kind of enforcement action, a Notice of Intended Enforcement Action (NOIE), was analyzed.

Figure E-10 presents a set of lift analysis results obtained for enforcement action events after NRs. The same three NR subsets were used as predictors with a 14-day outcome window and a range of evidence window widths. Tests indicate that using Type 3 NRs yields significant lifts for 7-, 14- and 28-day outcome windows, equaling 1.4, 1.37, and 1.3, respectively. Using all NRs as predictors of upcoming enforcement actions yields lifts of 1.18 and 1.2 for outcome windows of 7 and 14 days, respectively. Randomization tests were then performed using the bootstrapping method to obtain the confidence interval. In this case, the lower bound of the 95 percent confidence interval for these values was found to be slightly under 1.0. This may indicate less than desired robustness of the results for randomized choice of the sample subsets of establishments. (For a detailed description of the randomization procedure, refer to "Testing Significance of the Lift Statistic and AUC Scores," in the section titled "Overview of Analytic Methodology," later in this appendix.) Interestingly, the Industry Coalition defined NRs do not produce any significant correlations with enforcement actions. The results for Type 3 NRs are summarized in **Table E-5**.

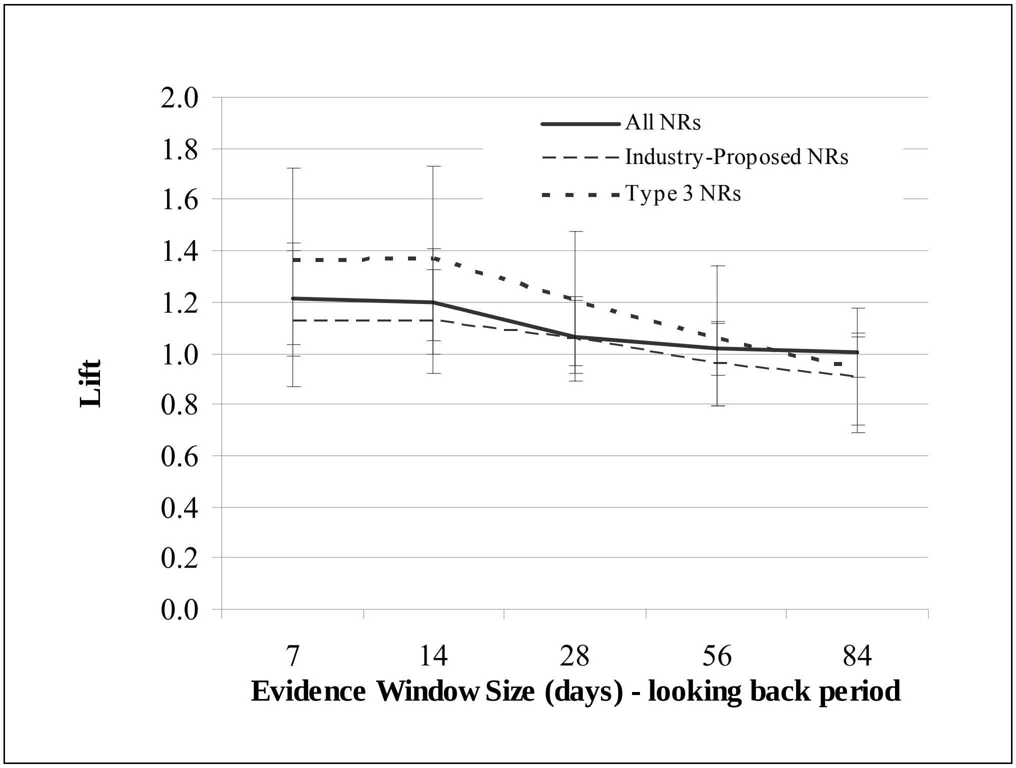


Figure E-10. Lift Analysis Results for NRs Versus NOIEs; Outcome Window Size is 14 Days

Table E-5. Relationship Between Type 3 NR Results and NOIE Enforcement Actions

Windows, Days			95% rCI		
Evidence	Outcome	Lift	Lower	Upper	p-value
7	7	1.2493	0.91102	1.61335	0.145
14	7	1.3937	1.08143	1.72962	**0.027**
28	7	1.2213	0.96563	1.50528	0.101
56	7	1.1013	0.83924	1.36818	0.256
84	7	0.9861	0.75834	1.21697	0.558
7	14	1.3615	1.03353	1.71982	**0.046**
14	14	1.369	1.05188	1.72732	**0.033**
28	14	1.2031	0.94854	1.47713	0.1

Windows, Days		Lift	95% rCI		p-value
Evidence	Outcome		Lower	Upper	
56	14	1.0547	0.79528	1.34418	0.35
84	14	0.9458	0.68898	1.17292	0.658
7	28	1.3288	1.00964	1.6913	0.053
14	28	1.3063	1.03194	1.62706	**0.034**
28	28	1.1222	0.8888	1.37883	0.227
56	28	0.9423	0.68006	1.20944	0.65
84	28	0.962	0.705	1.19661	0.585

NRs and RTE *L. monocytogenes* Alternatives

The 2003 FSIS *L. monocytogenes* Risk Assessment illustrates that certain control measures are effective in controlling *L. monocytogenes*. On the basis of those control measures, establishments producing post-lethality exposed RTE meat and poultry products under FSIS jurisdiction choose one of several options, called Alternatives, to control *L. monocytogenes*. The *L. monocytogenes* Alternatives are:

- Alternative 1: Application of a post-lethality treatment to the RTE product to reduce or eliminate microorganisms on product and the use of an antimicrobial agent or process as part of the product formulation.
- Alternative 2a: Post-lethality treatment to limit the growth of *L. monocytogenes* on the product.
- Alternative 2b: Use of an antimicrobial agent or process as part of the product formulation.
- Alternative 3: Reliance on testing and sanitation measures only.

The FSIS has conducted analyses of subsets of NRs to see if there is any correlation between the number of NRs issued and voluntary adoption of post-lethality processing, antimicrobial agents, and/or sanitation procedures (i.e., *L. monocytogenes* Alternatives 1 through 3). In this case, we are examining the establishment's choice of *L. monocytogenes* control measure as a potential predictor of PBIS noncompliances (NRs) rather than using the NRs as a predictor (as was done in the other analyses).

The alternative control data was collected as a one-time set of data in September 2006; therefore, the NR data was examined from the PBIS datasets following this date. In this analysis, two subsets of PBIS data are considered: one covering 6 months starting in October 2006, and the other using only the month of October 2006. The analyses have been performed against the three subsets of NRs (all NRs, Industry Coalition definition of NRs relevant to public health, and FSIS Type 3 NRs), for four groups of establishments which use specific control Alternatives 1, 2a, 2b, and 3 in order of strictness, as well as for all considered establishments, irrespective of any control alternatives.

Tables E-6 and E-7 summarize the results. The first column contains the type of *Lm* Alternative control measure chosen by the establishment. The second column contains the number of establishments in each subset. The third column provides the average frequency of NR citations per day per establishment. The fourth column provides the randomization test result (denoted by +/– sign where appropriate) for significance of the difference of NR frequency between a specific subset of establishments versus all establishments. Lift 1 in the fifth column is calculated simply as the ratio of the NR frequency of specific subset of establishments to the average frequency for all considered establishments. The sixth column provides the percentage of establishments recording at least one of the specific types of NR over the period of analysis. The seventh column provides the randomization test result on this measure. Lift 2 in the eighth column is derived in a similar manner as Lift 1. Entries that are significantly higher than expected (at the confidence level of 95 percent) are marked with "+;" those that are significantly lower than expected are marked with "–."

Table E-6 presents the results obtained using PBIS NR data ranging from October 2006 through March 2007. Table E-7 covers the month of October 2006.

An interesting observation from these tables is that the proportion of establishments with NR occurrences reported over the period of observation is consistently higher among the establishments that apply more strict alternative control measures, and this trend applies to all three subsets of NRs.

Table E-6. Relationship Between NRs and RTE *L. monocytogenes* Alternative (October 2006 through March 2007)

L. monocytogenes **Alternative**	**Number of Est.**	**No. of NRs per Day**	**Sig**	**Lift 1**	**Est. with at Least One NR, %**	**Sig**	**Lift 2**
All NRs							
Alternative 1	203	0.0574	+	1.390	88.6700		1.013
Alternative 2a	654	0.0541	+	1.310	90.2141	+	1.031
Alternative 2b	72	0.0331		0.801	87.5000		1.000
Alternative 3	1,371	0.0332	–	0.805	86.0686		0.983
All Establishments	2,300	0.0413			87.5217		
Industry-proposed NRs							
Alternative 1	203	0.0380	+	1.519	77.3399		1.054
Alternative 2a	654	0.0350	+	1.400	77.2171	+	1.053
Alternative 2b	72	0.0192		0.766	73.6111		1.004
Alternative 3	1,371	0.0186	–	0.745	70.8972		0.967
All Establishments	2,300	0.0250			73.3478		
Type 3 NRs							
Alternative 1	203	0.0186	+	1.785	60.5911	+	1.263
Alternative 2a	654	0.0157	+	1.503	55.8104	+	1.164
Alternative 2b	72	0.0095		0.913	47.2222		0.985
Alternative 3	1,371	0.0068	–	0.649	42.3778	–	0.884
All Establishments	2,300	0.0104			47.9565		

Notes: + denotes results significantly higher than expected (at 95 percent confidence level, based on randomization test).
– denotes results significantly lower than expected (at 95 percent confidence level, based on randomization test).
Lift 1=average number of NRs per day for specific subset of establishments divided by the average number of NRs per day computed for all establishments.
Lift 2=percentage of establishments with at least one NRs for specific subset of establishments divided by the analogical percentage computed for all establishments.

Table E-7. Relationship Between NRs and RTE *L. monocytogenes* Alternative (October 2006)

L. monocytogenes Alternatives	Number of Est.	No. of NRs per Day	Sig	Lift 1	Est. with at Least One NR, %	Sig	Lift 2
All NRs							
Alternative 1	203	0.0635	+	1.393	57.6355		1.054
Alternative 2a	654	0.0617	+	1.352	61.3150	+	1.121
Alternative 2b	72	0.0377		0.827	52.7778		0.965
Alternative 3	1,371	0.0357	–	0.783	51.2035	–	0.936
All Establishments	2,300	0.0456			54.6957		
Industry-proposed NRs							
Alternative 1	203	0.0431	+	1.610	45.8128	+	1.243
Alternative 2a	654	0.0380	+	1.420	43.1193	+	1.170
Alternative 2b	72	0.0223		0.834	43.0556		1.168
Alternative 3	1,371	0.0192	–	0.718	32.2392	–	0.874
All Establishments	2,300	0.0268			36.8696		
Type 3 NRs							
Alternative 1	203	0.0216	+	1.824	23.6453	+	1.366
Alternative 2a	654	0.0182	+	1.537	24.4648	+	1.414
Alternative 2b	72	0.0114		0.962	19.4444		1.124
Alternative 3	1,371	0.0074	–	0.624	12.8374	–	0.742
All Establishments	2,300	0.0119			17.3044		

Notes: + denotes results significantly higher than expected (at 95 percent confidence level, based on randomization test).
– denotes results significantly lower than expected (at 95 percent confidence level, based on randomization test).
Lift 1=average number of NRs per day for specific subset of establishments divided by the average number of NRs per day computed for all establishments.
Lift 2=percentage of establishments with at least one NRs for specific subset of establishments divided by the analogical percentage computed for all establishments.

Conclusion: NRs as a Component of Public-Health Risk-Based Inspection

In this section (and following sections), the presence of positive pathogen results within an establishment has been used as a proxy for measuring loss of process control. The positive pathogen results for *Salmonella* are far more numerous than those for other pathogens and have therefore provided a much more robust statistical measure. It appears from these results that NRs can serve as a useful tool for anticipating problems within establishments. The lift results show that the Type 3 group of NRs is particularly good at predicting *Salmonella* problems. In other cases, the Industry Coalition group was the better indicator of future problems. The weakness of the All NR group as a predictor is probably due to the inclusion of many noncleanliness-related items, as was pointed out in the criticism of the original RBI algorithm, that is, items not as directly linked to public health.

The breadth of the NR dataset and its close relationship to establishment process control (once the noncleanliness NRs are filtered out) makes it a strong candidate for inclusion as a component of RBI. These analyses show that NRs should be included in any future RBI algorithms; however, the filtering of NRs to define the optimum predictors may require further work.

FOOD SAFETY CONSUMER COMPLAINTS

As discussed in Appendix D, some consumer complaints could be an indication of an establishment's ability to maintain an effective food safety system. In this section, analyses are presented that examine the relationship between food-safety-related consumer complaints and other indicators of food safety system performance. Specifically, analyses have been conducted to evaluate if there is a subset of consumer complaints that can be linked to other indicators of an establishment's food safety performance. To do that, a subset of consumer complaints was compared against pathogen test results, recalls, enforcement actions, and, for some consumer complaints, *L. monocytogenes* Alternatives. The analysis addresses two separate definitions of complaints considered relevant: OPEER and EPI. The relationship between NRs and consumer complaints was examined above, and they were found to be only marginally related.

Consumer Complaints and Pathogen Test Results

Analyses were conducted to find a possible correlation between public-health-related food safety consumer complaints and food safety performance as measured by pathogen (i.e., *Salmonella, L. monocytogenes, and E. coli* O157:H7*)* test results, for applicable product types. The analysis did not yield indications of significant correlations between pathogen data and consumer complaint data. The most significant finding generated a lift of 1.57 for the relationship between CCMS OPEER cases and M2K *Salmonella* positives, in which both evidence and outcome window widths were set to 7 days (p-value of 0.087). However, the upper and lower randomization 95 percent confidence levels on that value of lift were very wide (0.17 and 2.95, respectively) making the model unreliable for practical purposes.

Consumer Complaints and Food Safety Recalls

A food safety recall may be triggered by a variety of factors once the product has entered commerce. The recall is classified based upon the relative health risk, and a Class I recall is a situation where the product has a *reasonable* probability of causing a health risk if eaten. Analyses of a subset of food safety consumer complaints as they correlate to Class I recalls would assess whether there is a relationship between the two parameters, and whether consumer complaint history might be predictive of an establishment's recall history. However, the currently available supply of data does not allow for meaningful analyses because during the period of time under consideration (April 2006 to September 2006), there are only three establishments that appear in both the CCMS OPEER cut and in the recall.

Consumer Complaints and Enforcement Actions

Enforcement actions are an indicator of an establishment's performance and may also be considered to measure the efficacy of the food safety system. Analyses of a subset of food safety consumer complaints as they correlate to enforcement actions may indicate whether consumer complaints might be a predictor of an establishment's food safety system design. Again, the limited supply of relevant data prevented such analyses.

Between April 2006 and September 2006 there are no establishments listed in both the CCMS OPEER cut and in the enforcement actions datasets.

Consumer Complaints and RTE *L. monocytogenes* Alternative

As with the NR data, FSIS has conducted analyses of a subset of consumer complaints (CCMS data) presumed to be potentially related to *L. monocytogenes* to see if there is any correlation between the number of consumer complaints issued and voluntary adoption of post-lethality processing, antimicrobial agents, and/or sanitation procedures (i.e., *L. monocytogenes* Alternatives 1 through 3). These results were generated with a similar methodology to that described in the section about correlations between NRs and *L. monocytogenes* control alternatives (see "NRs and RTE *L. monocytogenes* Alternatives" section). In this case, we are examining the establishment's choice of *L. monocytogenes* control measures as a potential predictor of consumer complaints (as we did with NRs) rather than using the complaints as a predictor (as was done in the other analyses). **Table E-8** summarizes the results of analyzing the *L. monocytogenes* Alternative as a predictor of CCMS events. This analysis was obtained by using CCMS data (OPEER cut and EPI cut) from April 2006 to September 2006. Ideally, we would have chosen datasets that immediately follow the establishment's control measure report date (September 2006); however, this data was not available. For this analysis, we have assumed that the control measures were in place prior to the reporting date.

Table E-8. Relationship Between CCMS Data from OPEER and EPI Cut (from April to September 2006)

***L. monocytogenes* Alternatives**	**No. of Est.**	**No. of Consumer Complains per Day**	**Sig**	**Lift 1**	**Est. with at Least One Consumer Complaint, %**	**Sig**	**Lift 2**
OPEER							
Alternative 1	212	0.0006		1.555	7.0755		1.513
Alternative 2a	694	0.0007	+	2.058	8.5014	+	1.818
Alternative 2b	80	0.0001		0.196	1.2500		0.267
Alternative 3	1,494	0.0002	–	0.473	2.7443	–	0.587
All Establishments	2,480	0.0004			4.6774		
EPI							
Alternative 1	212	0.0002	+	2.700	2.3585		2.437
Alternative 2a	694	0.0001		1.512	1.4409		1.489
Alternative 2b	80	0.0000		0.000	0.0000		0.000
Alternative 3	1,494	0.0000		0.575	0.6024		0.622
All Establishments	2,480	0.0001			0.9677		

Notes:+ denotes results significantly higher than expected (at 95 percent confidence level, based on randomization test).
– denotes results significantly lower than expected (at 95 percent confidence level, based on randomization test).
Lift 1=average number of consumer complains per day for specific subset of establishments divided by the average number of consumer complains per day computed for all establishments.
Lift 2=percentage of establishments with at least one consumer complains for specific subset of establishments divided by the analogical percentage computed for all establishments.

We can observe a negative correlation between *L. monocytogenes* control data and CCMS records. It seems that establishments implementing stricter controls are more likely to be associated with a higher frequency of consumer complaints. Several possible explanations include: there could be confounding factors linked to both *L. monocytogenes* control and CCMS data, which may lead to the apparent correlation, such as establishment size (larger establishments that implement stricter control may also record more consumer complains because of high volumes of production); CCMS data is known to be susceptible to under-reporting; and CCMS data is sparse and only 6 months of data were analyzed, so it may be nonrepresentative.

Conclusion: Consumer Complaints as a Component of Public-Health Risk-Based Inspection

In general, very little evidence of correlation involving CCMS data was found. That can be attributed to the extreme sparseness of the CCMS data. The OPEER cut consisted of 423 cases in total collected over the period of April through September 2006; however, only 283 of these complaints could be matched to specific establishments. Since some establishments received multiple complaints, there were only 163 unique establishments associated with those cases. In the case of the EPI cut, out of 47 total complaints, 44 could be matched to one of 35 establishments. Such low volumes of data make it very unlikely for the currently used analytic methodology to spot relationships that deviate significantly from random chance. As more data is collected it may be possible to demonstrate a statistical relationship between consumer complaints and a loss of process control.

Even though such a relationship has yet to be demonstrated statistically, it is logical that consumer complaints (once filtered by the cut events) are related to process. The presence of complaints against an establishment could therefore be included in an RBI algorithm as one component of a larger "compliance measure." As more data is collected, the proper weighting of consumer complaints within this measure can be reevaluated.

FOOD SAFETY RECALLS

As discussed in Appendix D, a food safety recall is a voluntary action by a manufacturer or distributor of a meat or poultry product to protect the public from products that may cause health problems or possible death. Analyses were conducted on the correlation between food safety recalls and other potential indicators of food safety system performance. In each case the presence or absence of a previous recall was examined as a potential predictor of the other indicators. The results for the analyses between recalls and pathogen test results, enforcement actions, and RTE *L. monocytogenes* Alternative are discussed below. Results of analyses examining the relationships with the other parameters (NRs and consumer complaints) have already been discussed in the previous sections.

When the U. S. Department of Agriculture (USDA) Recall Committee recommends a recall, they classify the recall into one of three classes based on the relative health risk:

- Class I recalls are the most serious and involve a health hazard situation in which there is a reasonable probability that eating the food will cause health problems or death.
- Class II recalls involve a potential health hazard situation in which there is a remote probability of adverse health consequences from eating the food.
- Class III recalls involve a situation in which eating the food will not cause adverse health consequences.

The data used in the analyses cover a 3-year period from March 2004 through March 2007, and are rather sparse. The dataset consists of 135 recalls, including 132 which could be associated with one of 120 unique establishments. Ten of the establishments recorded more than one recall. There are 113 of Class I recalls, 12 of Class II, and 7 of Class 3. The analyses have been conducted using two groupings of recalls: a set of all recalls, and a set excluding Class 3 recalls (i.e., excluding the recalls not likely to cause health consequences). Given the very small number of Class III recalls, the results of analyses are not significantly different between these sets.

Recalls and Pathogen Test Results

Analyses have been conducted to examine the correlation of public-health-related food safety recalls with food safety performance as measured by pathogen (i.e., *Salmonella, L. monocytogenes,* and *E. coli* O157:H7*)* test results, for applicable product types. Most of the results of those analyses turned out to be statistically insignificant. However, some statistical significance is associated with the correlations between *L. monocytogenes p*athogen test results and the food safety recalls (Class I and Class II). It is likely that these results could be explained by the fact that over one third of the recall cases are actually related to *L. monocytogenes* contamination (for specific numbers, see the section titled "Overview of Data Sources," in this appendix).

Figure E-11 presents lift for the 28-day outcome window width. This outcome window width produced the best results from among those tested. The graphs computed for the two sets of recall classes are practically identical. The highest lift is observed at the 28-day evidence window width and its value slightly exceeds 10.0 at the p-value of randomization test of significance of 0.001. Its randomization confidence interval appears to be relatively wide. The results for shorter evidence window widths are not significant with lower lifts, while those for longer windows also correspond to lower lifts. The relatively high lifts are not seconded by convincing AUC scores for they are very close to 0.5.

Recalls and Enforcement Actions

Analyses of a subset of food safety recalls to assess if they are correlated with enforcement actions were also performed. The results of such analyses for the two recall subsets (set of all recalls and set of Class I and II recalls) as predictors of enforcement actions, using a 56-day outcome window width, are shown in **Figure E-12**. This outcome window width produced the best results among those tested. The lift series for the set of all recalls and the set of Class I and II practically overlap, which indicates that Class III recalls have essentially no effect on the analysis. Lifts computed for the evidence windows 7, 14, and 28 days wide have been found statistically significant; however, the observed bands between the upper and lower limits of 95 percent confidence intervals obtained from randomization test are relatively wide.

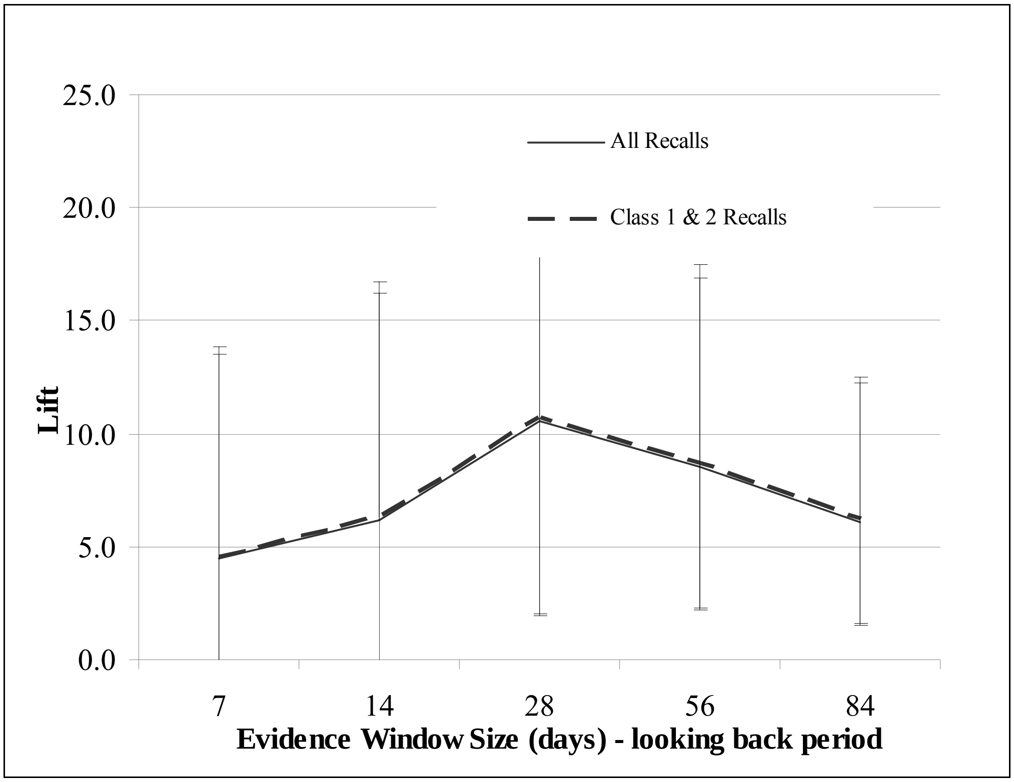


Figure E-11. Lift for the Relationship Between Recalls and *L. monocytogenes* Pathogen Test Results; Outcome Window Size is 28 Days

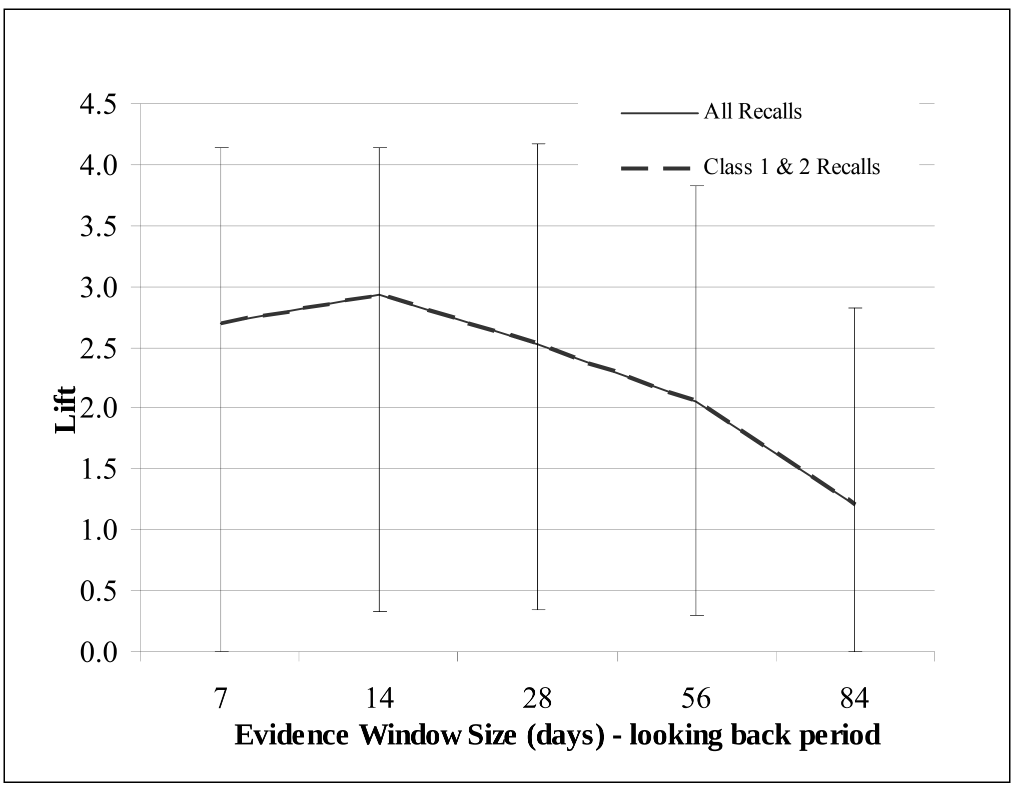


Figure E-12. Lift Results for the Relationship Between Recalls and Enforcement Actions

The results indicate that using recall information gathered over the last 56 days (both for Class I and II, as well as for all recalls) may be useful for predicting enforcement actions in the following 7 and 14 days, as it yields significant lifts of 3.16 and 3.39, respectively, with p-values of 0.013 and 0.01. The upper and lower limits of 95 percent confidence interval obtained by randomization test are within reasonable ranges (from 1.17 to 5.68 for 7-day outcome window and from 1.44 to 6.37 for 14-day outcome window width). **Table E-9** details these results.

Table E-9. Lift Statistics for Enforcement Action after Recalls from March 2004 to March 2007 for Meat and Poultry Product

Recall Classes	Windows, Days		Lift	95% rCI		p-value
	Evidence	Outcome		Lower	Upper	
1 and 2*	7	7	0.668461	0	2.122281	0.298
1 and 2	14	7	2.409138	0	7.69488	0.122
1 and 2	28	7	1.291416	0	4.296231	0.307
1 and 2	56	7	3.15521	1.169368	5.677783	**0.013**
1 and 2	84	7	1.864044	0.206446	2.992203	0.12
1 and 2	7	14	2.39052	0	7.261506	0.1
1 and 2	14	14	2.24772	0	8.084992	0.137
1 and 2	28	14	1.503244	0	4.261596	0.223
1 and 2	56	14	3.393194	1.440781	6.372998	**0.01**
1 and 2	84	14	1.995854	0.161209	3.288538	0.095
All (1, 2, and 3)	7	7	0.668461	0	2.122281	0.298
All (1, 2, and 3)	14	7	2.409138	0	7.69488	0.122
All (1, 2, and 3)	28	7	1.291416	0	4.296231	0.307
All (1, 2, and 3)	56	7	3.15521	1.169368	5.677783	**0.013**
All (1, 2, and 3)	84	7	1.864044	0.206446	2.992203	0.12
All (1, 2, and 3)	7	14	2.39052	0	7.261506	0.1
All (1, 2, and 3)	14	14	2.24772	0	8.084992	0.137
All (1, 2, and 3)	28	14	1.503244	0	4.261596	0.223
All (1, 2, and 3)	56	14	3.393194	1.440781	6.372998	**0.01**
All (1, 2, and 3)	84	14	1.995854	0.161209	3.288538	0.095

* Union of Class 1 and Class 2 recalls.

Recalls and RTE *L. monocytogenes* Alternative

FSIS has conducted analyses of recalls thought to be potentially related to *L. monocytogenes* to see if there is any correlation between the number of recalls issued and voluntary adoption of post-lethality processing, antimicrobial agents, and/or sanitation procedures (i.e., *Lm* Alternatives 1 through 3). Similar analysis to that explained in the section addressing relationships between NRs and RTE *L. monocytogenes* Alternative control (see "NRs and RTE *Lm* Alternatives" section) has been applied here. **Table E-10** summarizes the results of examining the relationship between recall data ranging from April 2006 through September 2006 and RTE *L. monocytogenes* Alternative control data. A negative correlation pattern similar to that discussed above in the context of CCMS versus alternative control can be seen here as well. As explained previously, this could be attributable to the sparseness of recall data and to the existence of confounding factors.

Table E-10 Relationship Between *L. monocytogenes* Alternatives and Recalls from April to September 2006

L. monocytogenes Alternatives	Number of Est.	No. of Recalls per Day	Sig	Lift 1	Est. with at Least One Recall, %	Sig	Lift 2
All Recalls							
Alternative 1	212	0.0003		1.712	3.3019		1.137
Alternative 2a	694	0.0002		1.307	3.7464		1.290
Alternative 2b	80	0.0001		0.378	1.2500		0.431
Alternative 3	1,494	0.0001		0.789	2.5435		0.876
All Establishments	2,480	0.0002			2.9032		
Class I & II Recalls							
Alternative 1	212	0.0003		1.650	2.8302		1.017
Alternative 2a	694	0.0002		1.283	3.6023		1.295
Alternative 2b	80	0.0001		0.397	1.2500		0.449
Alternative 3	1,494	0.0001		0.809	2.4766		0.890
All Establishments	2,480	0.0002			2.7823		

Notes: + denotes results significantly higher than expected (at 95 percent confidence level, based on randomization test).
– denotes results significantly lower than expected (at 95 percent confidence level, based on randomization test).
Lift 1=average number of recalls per day for specific subset of establishments divided by the average number of recalls per day computed for all establishments.
Lift 2=percentage of establishments with at least one recall for specific subset of establishments divided by the analogical percentage computed for all establishments.

Conclusion: Food Safety Recalls as a Component of Public-Health Risk-Based Inspection

The presence of a recall indicates unequivocally that an establishment has lost process control at some point. For this reason alone, it is logical to include this information in an RBI algorithm. These analyses show that Class I and Class II recalls have a statistical relationship with *L. monocytogenes* contamination and might also serve as a predictor of future enforcement actions. The presence of previous recalls associated with an establishment can be included in an RBI algorithm as one component of a "compliance measure."

ENFORCEMENT ACTIONS

As discussed in Appendix D, there are a variety of enforcement actions the Agency can take against establishments that fail to sufficiently comply with applicable requirements—both food safety and non-food safety. For the previously proposed RBI algorithm, enforcement actions were given different weights depending on their severity. Analyses are described below that examine whether enforcement actions can be linked to other indicators of an establishment's food safety performance. To do that, a subset of enforcement actions was compared against pathogen test results, and, for some establishments that make RTE products, *L. monocytogenes* Alternative. A description of the enforcement action dataset is provided in the section titled "Overview of Data Sources." The relationship between enforcement actions and other parameters has been examined in the previous sections.

Enforcement Actions and Pathogen Test Results

Analyses have been conducted to examine the correlation of enforcement actions with food safety performance as measured by pathogen (i.e., *Salmonella*, *L. monocytogenes,* and *E. coli* O157:H7*)* test results, as they are applicable to product type. The results include a few combinations of evidence and outcome window widths which lead to significant p-value and computed lift greater than 1.0; however, 95 percent confidence intervals obtained are quite wide. This may be attributed to the sparseness of enforcement action data since most establishments have not been subjected to such actions during the period under analysis.

Table E-11 summarizes the results. Significant lifts are found when using enforcement action information collected over the last 84 days to predict *E. coli* positives over the next 28 or 56 days. This is also true using enforcement action records over the last 28, 56, and 84 days to predict positive *E. coli* tests over the outcome window of 84 days; however, the 95 percent confidence interval obtained from bootstrapping is too wide for that result to be considered reliable. Significant lift can also be observed when using records of enforcement actions over the last 28 days to predict *Salmonella* positives over the next 7 days, as well as using enforcement actions over the last 56 days to predict *Salmonella* positives over the next 56 days. Most of the results obtained using the 84-day outcome window also produce significant p-values. Unfortunately, the 95 percent confidence intervals from bootstrapping are quite wide although they are slightly narrower than in the case of *E. coli* analysis.

Table E-11. Correlation of Enforcement Actions with *E. coli*- and *Salmonella*-Positive Results, April through September 2006

Pathogen	Windows, Days		Lift	95% rCI		p-value
	Evidence	Outcome		Lower	Upper	
E. coli	7	28	0	0	0	1
E. coli	14	28	0	0	0	1
E. coli	28	28	0	0	0	1
E. coli	56	28	0	0	0	1
E. coli	84	28	17.317	0	54.5375	**0.035**
E. coli	7	56	0	0	0	1
E. coli	14	56	0	0	0	1
E. coli	28	56	0	0	0	1
E. coli	56	56	16.138	0	53.2554	0.059
E. coli	84	56	27.555	0	92.7374	**0.018**
E. coli	7	84	3.8796	0	14.0618	0.107
E. coli	14	84	18.268	0	62.7238	0.05
E. coli	28	84	32.215	0	101.735	**0.033**
E. coli	56	84	41.002	0	123.975	**0.028**
E. coli	84	84	33.843	0	111.037	**0.018**
Salmonella	7	7	1.5195	0	5.12128	0.265
Salmonella	14	7	1.7895	0	5.19579	0.156
Salmonella	28	7	2.3775	0	5.60369	**0.011**
Salmonella	56	7	1.3117	0	3.69617	0.085
Salmonella	84	7	0.8969	0.08553	2.06952	0.321
Salmonella	7	56	1.0647	0	2.78804	0.409

Pathogen	Windows, Days		Lift	95% rCI		p-value
	Evidence	Outcome		Lower	Upper	
Salmonella	14	56	1.2094	0	2.78294	0.188
Salmonella	28	56	1.2415	0	2.8312	0.125
Salmonella	56	56	1.5858	0.21853	3.24167	**0.024**
Salmonella	84	56	1.2808	0.0896	2.8181	0.17
Salmonella	7	84	2.0862	0.41987	3.93517	**0.018**
Salmonella	14	84	2.3829	0.67482	4.33135	**0.001**
Salmonella	28	84	2.5114	0.65671	4.52981	**0.002**
Salmonella	56	84	2.1334	0.43608	4.07052	**0.011**
Salmonella	84	84	1.9435	0.35085	3.64448	0.06

Enforcement Actions and RTE *L. monocytogenes* Alternatives

Analyses were performed to see if there was any correlation between the voluntary adoption of post-lethality processing, antimicrobial agents, and/or sanitation procedures (i.e., *L. monocytogenes* Alternatives 1 through 3) and enforcement actions thought to be potentially related to *L. monocytogenes*. This required similar analysis as for NR versus *L. monocytogenes* controls (see "NRs and RTE *Lm* Alternatives" section). The results based on the enforcement action occurrence during the period from April 2006 to September 2006 are summarized in **Table E-12**. The frequency of actions for establishments that implement control Alternative 1 and those implementing Alternative 2a are comparable. Establishments that implement Alternative 3 seem to be more likely to get enforcement actions than others. These results should be taken with caution given the limited amount of available evidence and limited supply of enforcement actions data.

Table E-12 Relationship Between *L. monocytogenes* Alternatives and Enforcement Action (NOIE) Occurrences from April to September 2006

L. monocytogenes Alternatives	Number of Est.	No. of Enforcement Actions per Day	Sig	Lift 1	Est. with at Least One Enforcement Action, %	Sig	Lift 2
Alternative 1	212	0.0001		0.731	0.9434		0.731
Alternative 2a	694	0.0000		0.558	0.7205		0.558
Alternative 2b	80	0.0000		0.000	0.0000		0.000
Alternative 3	1,494	0.0001		1.297	1.6734		1.297
All Establishments	2,480	0.0001			1.2903		

Notes: + denotes results significantly higher than expected (at 95 percent confidence level, based on randomization test).
– denotes results significantly lower than expected (at 95 percent confidence level, based on randomization test).
Lift 1=average number of enforcement actions per day for specific subset of establishments divided by the average number of enforcement actions per day computed for all establishments.
Lift 2=percentage of establishments with at least one enforcement action for specific subset of establishments divided by the analogical percentage computed for all establishments.

Conclusion: Enforcement Actions as a Component of Public-Health Risk-Based Inspection

The sparseness of enforcement action data makes the analysis of it as a public health risk-based inspection component difficult. Lift calculations do show some predictive ability; however, the confidence intervals are quite wide. It is therefore not possible to justify statistically the presence of previous enforcement actions as a primary component of an RBI algorithm. However, because enforcement actions, by definition, indicate a loss of process control, they should still be considered for potential use as a component within an overall "compliance measure."

L. MONOCYTOGENES ALTERNATIVE CONTROL PROCESSES

As discussed in Appendix D, establishments that produce RTE products that are exposed to the environment subsequent to the lethality step must comply with the provisions of 9 CFR 430. The Agency maintains data that indicates how an establishment complies with those provisions, and therefore, how well they control the risk associated with *L. monocytogenes* in RTE products. The RTE *L. monocytogenes* Alternatives were taken into account in the RCM portion of the RBI algorithm proposed in Spring 2006, and were given different weights based upon which RTE Regulatory Alternative category an establishment would fall into. Analyses of possible correlations between *L. monocytogenes* Alternative control processes and *L. monocytogenes* test results for the applicable products are presented in this section.

The raw *L. monocytogenes* Alternative control information available for analysis involves 2,480 establishments which reported their control status as of September 2006. This was a one-time survey of plants, so the dataset is static (a single point in time) and self-reported. There are four distinct control states (in the decreasing level of control: Alternatives 1, 2a, 2b, and 3) and three control methods reported (sanitation, antimicrobial, and post-lethality). The lowest control state, alt3, implemented in 1,494 establishments, requires only that the sanitation method is implemented. Alternative 2b (80 establishments) requires sanitation and post-lethality; Alternative 2a (694 establishments) requires sanitation and antimicrobial measures, while Alternative 1 (212 establishments) requires implementation of all three control methods. In the raw data an additional category was encountered: Alternative 2. Since this category was not an official one it was assumed that Alternative 2 equates to Alternative 2a (this correction affected 48 establishments).

Since the alternative control information is static, the analysis was conducted using two overlapping periods of coverage of the microbial test data (M2K): from January 2005 to March 2007 and from October 2006 through March 2007. The analyses include establishments with known alternative control information and which have a record of at least one *L. monocytogenes* test conducted within the period of time considered.
Table E-13 summarizes the results.

Table E-13 presents three statistics intended to characterize the frequency of occurrences of positive *L. monocytogenes* tests. *L. monocytogenes* prevalence is defined as the mean ratio of the number of positive results to the total number of *L. monocytogenes* tests conducted, averaged across all considered establishments. The average number of *L. monocytogenes* positives per day is defined as the mean of the ratio of positive counts to the number of days within the period of analysis, averaged across all establishments. The likelihood of having at least one positive is defined as the mean proportion of establishments having at least one *L. monocytogenes* positive over the period of analysis. The extent of departure of the value of the individual statistic computed for a subset of establishments in a particular control state, from the expectation based upon all considered establishments, is measured by lift. Here lift is defined as the ratio of each statistic for an "alternative" to "All." The table also includes results of randomization tests of significance. The entry is marked with a "+" or "–" sign in the "sig" column if the relevant measure is significantly higher or lower than expected at the confidence level of 95 percent.

In this case the term "lift" is used in a slightly different context than before. It has the same practical meaning though, in that it measures the extent of departure of some statistic computed for a subset of data from its value computed for the baseline (usually the whole set of) data. The table above summarizes results obtained for three different statistics. These base statistics include prevalence and frequency of positives per day which are not binarized. Certain kinds of binarization are however involved in the third of the base statistics, where the proportion of establishments with any *L. monocytogenes* positives is examined. In this case the establishments are split into two classes: those without any *L. monocytogenes* issues, and all others. This binarization step is not present in the previous analyses.

Table E-13 Relationship Between *L. monocytogenes* Positives and *L. monocytogenes* Alternative Control Processes

Lm Control Alternatives	No. of Est.	Lm Prevalence	Lift	Sig	No. of Lm positives per day	Lift	Sig	Est. with at Least One Lm Positive, %	Lift	Sig
Using all Lm data from January 2005 through March 2007										
Alternative 1	185	0.013%	0.052	–	0.0000	0.266	–	0.68	0.413	–
Alternative 2a	654	0.207%	0.800		0.0001	0.904		1.55	0.935	
Alternative 2b	69	0.000%	0.000		0.0000	0.000		0.00	0.000	
Alternative 3	1,380	0.333%	1.288		0.0002	1.206		1.94	1.170	
All Establish-ments	2,288	0.258%			0.0001			1.66		
Using Lm data from October 2006 through March 2007										
Alternative 1	146	0.178%	0.335		0.0001	0.556		4.86	0.687	
Alternative 2a	516	0.450%	0.846		0.0001	0.956		6.73	0.950	
Alternative 2b	56	0.459%	0.863		0.0001	0.918		4.35	0.614	
Alternative 3	1,031	0.622%	1.169		0.0002	1.084		7.68	1.085	
All Establish-ments	1,749	0.532%			0.0002			7.08		

Notes: + denotes results significantly higher than expected (at 95 percent confidence level, based on randomization test).
– denotes results significantly lower than expected (at 95 percent confidence level, based on randomization test).
Lift 1=average number of enforcement actions per day for specific subset of establishments divided by the average number of enforcement actions per day computed for all establishments.
Lift 2=percentage of establishments with at least one enforcement action for specific subset of establishments divided by the analogical percentage computed for all establishments.
Notes: + denotes results significantly higher than expected (at 95 percent confidence level, based on randomization test).
– denotes results significantly lower than expected (at 95 percent confidence level, based on randomization test).
Absence of any sig designation means the result are not significantly different from expected (at 95 percent confidence level, based on randomization test).

It can be observed that all of the obtained results are not significant, except for the Alternative 1 control evaluated with *L. monocytogenes* prevalence rates over the whole set of the available data. This effect disappears when looking at the second set of data, which are collected after September 2006 (a shorter and more recent period of time shown in the bottom part of the table). Even though the obtained results are mostly insignificant, they follow an intuitive pattern that the stricter alternatives are related to the lower *L. monocytogenes* positives. For instance the prevalence of *L. monocytogenes* positives in establishments implementing alt1 control is only about 5 percent of the baseline measure taken across all of the considered establishments, while the prevalence for Alternative 3 establishments amounts to 129 percent of the baseline.

Table E-14 summarizes the results of randomization tests of significance for any observed differences in observed frequency of *L. monocytogenes* positives between all pairs of control states. The top part of the table presents the differences in prevalence rates, the middle shows p-values of the one-sided significance test for increase in prevalence, and the bottom part contains the p-values of the one-sided test of decrease in prevalence rate. The results correspond to the whole set of available M2K data: from January 2005 through March 2007. For this analysis it was assumed that whatever control measure was reported in September 2006 was in place for this whole period.

Table E-14. Randomization Test for *L. monocytogenes* Prevalence Rate Differences Among Alternatives (using all *L. monocytogenes* data)

***L. monocytogenes* Alternative**	***L. monocytogenes* Alternative**			
	Alternative 1	**Alternative 2a**	**Alternative 2b**	**Alternative 3**
Difference of Mean				
Alternative 1		−0.0027	−0.0028	−0.0044
Alternative 2a	0.0027		−0.0001	−0.0017
Alternative 2b	0.0028	0.0001		−0.0016
Alternative 3	0.0044	0.0017	0.0016	
P value				
Alternative 1		0.9596	0.8948	0.9910
Alternative 2a	0.0402		0.5622	0.8850
Alternative 2b	0.1106	0.4380		0.5914
Alternative 3	0.0118	0.1094	0.3962	
Neg P Value				
alt1		0.0370	0.1124	0.0098
alt2a	0.9674		0.4436	0.1168
alt2b	0.8992	0.5694		0.4138
alt3	0.9872	0.8840	0.5890	

The results indicate that establishments that implement Alternative 2a experience a significantly higher *L. monocytogenes* prevalence than those implementing Alternative 1, and those implementing Alternative 3 have significantly higher *L. monocytogenes* prevalence than those implementing Alternative 1. All other differences do not turn out to be significant. Analogous results obtained for two other statistics which could be used to measure difference in frequency in *L. monocytogenes* occurrences (average number of positives per day and the average proportion of establishments that report *L. monocytogenes* positives over the period of analysis) do not indicate significant differences between control states. Analogical results obtained for the most recent 6 months of M2K data include only one significant finding: the difference in the number of positives per day between establishments implementing Alternatives 2b and 3.

Table E-15 looks at the data from the point of view of the control method employed. Even though the number of establishments applying post-lethality measures is relatively small, they achieve a significant reduction in the *L. monocytogenes* prevalence and occurrence rates, with respect to the global averages.

The results of statistical tests of differences in the measurements have not been found to be significant. The one exception is that the post-lethality method has been found to be significantly more effective in terms of predicting the *L. monocytogenes* prevalence and the average number of the *L. monocytogenes* positives per day when compared against the observed performance of all establishments.

Table E-15. *L. monocytogenes* Prevalence and Occurrence Rates Relationship with *L. monocytogenes* Control Methods

Lm **Control Method**	**No. of Est.**	***Lm*** **Prevalence**	**Lift**	**Sig**	**No. of *Lm* positives per day**	**Lift**	**Sig**	**Est. with at Least One *Lm* Positive, %**	**Lift**	**Sig**
Using all Lm data from January 2005 until March 2007										
Anti-microbial	839	0.390%	0.733		0.0001	0.868		6.32	0.892	
Post-lethality	254	0.255%	0.478	–	0.0001	0.655		4.72	0.667	
All Establish-ments	2,288	0.532%			0.0002			7.08		
Using Lm data from October 2006 until March 2007										
Anti-microbial	662	0.164%	0.635		0.0001	0.763		1.36	0.820	
Post-lethality	202	0.010%	0.038	–	0.0000	0.192	–	0.50	0.299	
All Establish-ments	1,749	0.258%			0.0001			1.66		

Notes: + denotes results significantly higher than expected (at 95 percent confidence level, based on randomization test).
– denotes results significantly lower than expected (at 95 percent confidence level, based on randomization test).
Absence of any sig designation means the results are not significantly different from expected (at 95 percent confidence level, based on randomization test).

The available data contains some evidence of the effects of difference in the implemented *L. monocytogenes* Alternative control methods. However, given the scattered pattern of significant outcomes, it is difficult to draw general conclusions reaching beyond the intuitive (i.e., the stricter the control, the lower the likelihood of compromising public health).

Conclusion: *L. monocytogenes* Alternative as a Component of Public-Health Risk-Based Inspection

As previously mentioned, the data on *L. monocytogenes* Alternatives within establishments was taken from self-reported information in September 2006. The data is therefore static (one-time information for each responding establishment) and may contain several biases (only establishments with known problems may have chosen strong measures, only establishments without known problems may have responded, etc.). In addition, in order to perform the analysis assumptions had to be made as to when the control measures were put into place.

The analyses do not show that the choice of *L. monocytogenes* Alternative is a strong predictor for any of our measures of process control.

Other Potential Factors – Establishment Characteristics Collected in RTI Survey

In addition to those parameters used in the RBI algorithm presented previously, FSIS has been exploring other parameters that could be incorporated into an algorithm for use in directing resources. It is important that FSIS focus not only on the data previously used, but also other data that it has that could be used and data that could possibly be available to it for use in the future. This section presents the results of analyses evaluating some other potential data, as well as discussing what analyses should be considered in the future if other data becomes available.

As described in Appendix D, RTI International conducted a voluntary, OMB-approved survey of FSIS-regulated processing facilities to gather information on establishment characteristics, including age of production facility, production space square footage, number of employees, HACCP training, use of chemical sanitizers, and the number of inspectors. FSIS requested that RTI conduct a statistical analysis to determine whether any of those characteristics are related to the pathogen testing results (specifically, *Salmonella* and *Listeria* test results), and if they would be appropriate to use in an RBI algorithm. Such analyses are important to determine the potential usefulness of data on other establishment characteristics and to assess whether efforts should be made to acquire these data on an ongoing basis in the future.

The analysis focused on two types of processing establishments: those that produce ground beef and those that produce RTE meat and poultry products. The outcome measure used for the analysis is whether or not an establishment had one or more Salmonella test results (including *Listeria* test results in the case of RTE establishments) over the 2004 through 2006 period. Of the 108 ground beef establishments that responded to the voluntary survey, 57 establishments had 1 or more positive *Salmonella* test results. Of the 343 RTE establishments that responded to the voluntary survey, 35 had 1 or more positive *Salmonella* or *Listeria* test results.

The summary statistics were calculated on the differences in characteristics of establishments based on whether the establishment had one or more positive pathogen test results. The results for ground beef establishments are presented in **Table E-16**, and the results for RTE establishments are presented in **Table E-17**. Means and standard deviations are presented for continuous variables and frequencies, and percentages are presented for categorical variables. For ground beef establishments, variables that were significantly different at the 10 percent level included the percentage of time a food safety manager is dedicated to food safety activities, whether food safety training is provided to new employees, and the number of HACCP-trained employees. For RTE establishments, the only variable that was significantly different at 10 percent alpha level or better was the lot (or batch) size. Because the univariate analyses do not control for other establishment characteristics that affect performance, multivariate analyses were subsequently conducted using the complete set of variables available in the datasets.

Table E-16. Descriptive Statistics for Key Variables for Ground Beef Establishments

		No. of Positive *Salmonella* Tests (N = 51)		**One or More Positive *Salmonella* Tests (N = 57)**		**All Establishments (N = 108)**		
Q#	Voluntary Survey Question	Mean	Std	Mean	Std	Mean	Std	p-value
4.1	Calendar year plant was built or recently renovated.	1989	16	1991	15	1990	16	0.51
4.2	Approximate total square footage of the production space	54,850	104,415	45,766	98,025	50,055	100,719	0.64
4.8	Approximately how many people are employed at this plant?	170	383	131	268	150	326	0.55
		N	%	N	%	N	%	
4.10	Plant has a person on staff whose primary responsibility is to manage food safety activities at the plant.	39	76.5	36	63.2	75	69.4	0.13
4.11	Approximately what percentage of this plant's food safety manager's time is devoted to managing food safety activities at the plant?							
	0. 0 percent	12	23.5	21	36.8	33	30.6	0.10
	1. 1 to 24 percent	13	25.5	7	12.3	20	18.5	
	2. 25 to 49 percent	9	17.7	11	19.3	20	18.5	
	3. 50 to 74 percent	3	5.9	9	15.8	12	11.1	
	4. 75 to 99 percent	8	15.7	7	12.3	15	13.9	
	5. 100 percent	6	11.8	2	3.5	8	7.4	
4.12	This plant has a quality control/ quality assurance department.	27	52.9	35	61.4	62	57.4	0.37
		Mean	Std	Mean	Std	Mean	Std	
4.7	For the meat or poultry product with the highest production volume, what is the average lot size (pounds)?	28,009	85,031	18,107	33,647	22,783	63,213	0.44
	Number of inspectors (2005)	1.0	0.6	1.2	0.8	1.1	0.7	0.30
		N	%	N	%	N	%	
4.5	How many processing shifts does this plant usually operate per day?							
	1. One	40	78.4	36	63.2	76	70.4	0.13
	2. Two	11	21.6	19	33.3	30	27.8	
	3. Three	0	0.0	2	3.5	2	1.9	
4.16	What was the approximate value of total plant sales revenue for the most recently completed fiscal year?							
	1. Under $249,999	7	13.7	8	14.0	15	13.9	0.21
	2. $250,000 to $499,999	3	5.9	5	8.8	8	7.4	
	3. $500,000 to $1.49 million	8	15.7	5	8.8	13	12.0	
	4. $1.5 to $2.49 million	7	13.7	1	1.8	8	7.4	

		No. of Positive *Salmonella* Tests (N = 51)		**One or More Positive *Salmonella* Tests (N = 57)**		**All Establishments (N = 108)**		
	5. $2.5 to $24.9 million	13	25.5	20	35.1	33	30.6	
	6. $25 to $49.9 million	4	7.8	8	14.0	12	11.1	
	7. $50 to $99.9 million	4	7.8	5	8.8	9	8.3	
	8. $100 to $249.9 million	3	5.9	5	8.8	8	7.4	
	9. $250 to $499.9 million	2	3.9	0	0.0	2	1.9	
	10. $500 to $999.9 million	0	0.0	0	0.0	0	0.0	
	11. $1 billion or more	0	0.0	0	0.0	0	0.0	
3.1	Food safety training is provided for newly hired production employees of this plant.	15	29.4	8	14.0	23	21.3	0.05
3.2	Continuing food safety training is provided for production employees of this plant.	12	23.5	19	33.3	31	28.7	0.26
3.3	Approximately how many production and retail employees currently working at this plant have completed formal HACCP training?							
	1. None	10	19.6	6	10.5	16	14.8	0.02
	2. 1 to 3 employees	25	49.0	32	56.1	57	52.8	
	3. 4 to 9 employees	6	11.8	16	28.1	22	20.4	
	4. 10 to 20 employees	10	19.6	3	5.3	13	12.0	
	5. More than 20 employees	0	0.0	0	0.0	0	0.0	

Table E-17. Descriptive Statistics for Key Variables for RTE Establishments

		No. of Positive *Salmonella* or *Listeria* Tests (N = 308)		**One or More Positive *Salmonella* or *Listeria* Tests (N = 35)**		**All Establishments (N = 343)**		
#	Voluntary Survey Question	Mean	Std	Mean	Std	Mean	Std	p-value
4.1	Calendar year plant was built or recently renovated.	1990	16	1987	21	1989	17	0.47
4.2	Approximate total square footage of the production space	73,515	176,803	52,431	99,687	71,363	170,554	0.29
4.8	Approximately how many people are employed at this plant?	148	278	130	219	146	27	0.66
		N	%	N	%	N	%	
4.10	Plant has a person on staff whose primary responsibility is to manage food safety activities at the plant.	216	70.1	27	77.1	243	70.9	0.39
4.11	Approximately what percentage of this plant's food safety manager's time is devoted to managing food safety activities at the plant?							
	0. 0 percent	92	29.9	8	22.9	100	29.2	0.73
	1. 1 to 24 percent	56	18.2	7	20.0	63	18.4	

		No. of Positive *Salmonella* or *Listeria* Tests (N = 308)		**One or More Positive *Salmonella* or *Listeria* Tests (N = 35)**		**All Establishments (N = 343)**		
	2. 25 to 49 percent	41	13.3	5	14.3	46	13.4	
	3. 50 to 74 percent	43	14.0	8	22.9	51	14.9	
	4. 75 to 99 percent	46	14.9	5	14.3	51	14.9	
	5. 100 percent	30	9.7	2	5.7	32	9.3	
4.12	This plant have a quality control/quality assurance department.	198	64.3	22	62.9	220	64.1	0.87
		Mean	Std	Mean	Std	Mean	Std	
4.7	For the meat or poultry product with the highest production volume, what is the average lot size?	23,864	63,284	14,733	20,964	22,932	60,385	0.07
	Number of inspectors (2005)	1.1	0.8	0.9	0.8	1.1	0.8	0.18
		N	%	N	%	N	%	
4.5	How many processing shifts does this plant usually operate per day?							
	1. One	214	69.5	23	65.7	237	69.1	0.23
	2. Two	85	27.6	9	25.7	94	27.4	
	3. Three	9	2.9	3	8.6	12	3.5	
4.16	What was the approximate value of total plant sales revenue for the most recently completed fiscal year?							
	1. Under $249,999	29	9.4	3	8.6	32	9.3	0.33
	2. $250,000 to $499,999	26	8.4	1	2.9	27	7.9	
	3. $500,000 to $1.49 million	50	16.2	3	8.6	53	15.5	
	4. $1.5 to $2.49 million	29	9.4	5	14.3	34	9.9	
	5. $2.5 to $24.9 million	91	29.6	14	40.0	105	30.6	
	6. $25 to $49.9 million	21	6.8	2	5.7	23	6.7	
	7. $50 to $99.9 million	27	8.8	1	2.9	28	8.2	
	8. $100 to $249.9 million	21	6.8	4	11.4	25	7.3	
	9. $250 to $499.9 million	9	2.9	0	0.0	9	2.6	
	10. $500 to $999.9 million	5	1.6	2	5.7	7	2.0	
	11. $1 billion or more	0	0.0	0	0.0	0	0.0	
3.1	Food safety training is provided for newly hired production employees of this plant.	79	25.7	9	25.7	88	25.7	0.99
3.2	Continuing food safety training is provided for production employees of this plant.	91	29.6	12	34.3	103	30.0	0.56
3.3	Approximately how many production and retail employees currently working at this plant have completed formal HACCP training?							
	1. None	24	7.8	0	0.0	24	7.0	0.27
	2. 1 to 3 employees	184	59.7	25	71.4	209	60.9	
	3. 4 to 9 employees	61	19.8	6	17.1	67	19.5	
	4. 10 to 20 employees	23	7.5	1	2.9	24	7.0	
	5. More than 20 employees	16	5.2	3	8.6	19	5.5	

Further statistical analyses were conducted to determine which characteristics of establishments were associated with a statistically significant increase or decrease in the likelihood of one or more positive pathogen test results. Segmentation analysis (in this case, CART analysis) was conducted to identify which variables among the large number of variables in the datasets had an appreciable degree of explanatory power related to pathogen testing results. Because of the low number of positive test results for RTE establishments, the segmentation analysis was sufficient for identifying important variables that are associated with pathogen testing results. For ground beef establishments, factor analysis and logistic regressions were conducted to determine whether the results would provide additional information beyond that provided in the segmentation analysis.

Results of Analysis for Ground Beef Establishments

Figure E-13 shows the results of the segmentation analysis for ground beef establishments. Some 65 potential variables for ground beef establishments were included in the analysis. Among those variables, pounds of beef products produced emerged as the strongest predictor of establishment performance as measured by Salmonella test results. Specifically, among all establishments, the odds of passing (that is, having no positive Salmonella test results from 2004 through 2006) are over 3 times higher for those producing less than or equal to 250,000 pounds of beef products during the past year. As such, the 108 analyzed establishments are classified into two groups: 75 “lower volume” establishments on the left branch of the classification tree, and 33 “higher-volume” establishments on the right branch. For “higher-volume” establishments:

- The odds of passing are one-tenth for establishments with fewer than 9 production employees who have completed formal HACCP training as compared to establishments with more HACCP trained employees.
- Among the above establishments with fewer than 9 HACCP trained production employees, the odds of passing are 40 times higher when facility NR rate is less than 0.3 percent.

For “lower-volume” establishments:

- Among establishments with a facility NR rate over 11.6 percent, establishments are much less like to pass if they have smaller production spaces (less than or equal to 1,250 square feet) as compared to establishments with larger production spaces.
- Among establishments with a facility NR rate less than or equal to 11.6 percent, establishments with a sanitation NR rate less than or equal to 0.1 percent are almost 7 times more likely to pass. However, when the sanitation NR rate for such establishments is over 0.1 percent, the odds of passing are over 6 times higher when the establishment has a food safety manager on staff. Furthermore, the latter establishments are more likely to pass if their lot sizes are less than 800 pounds.

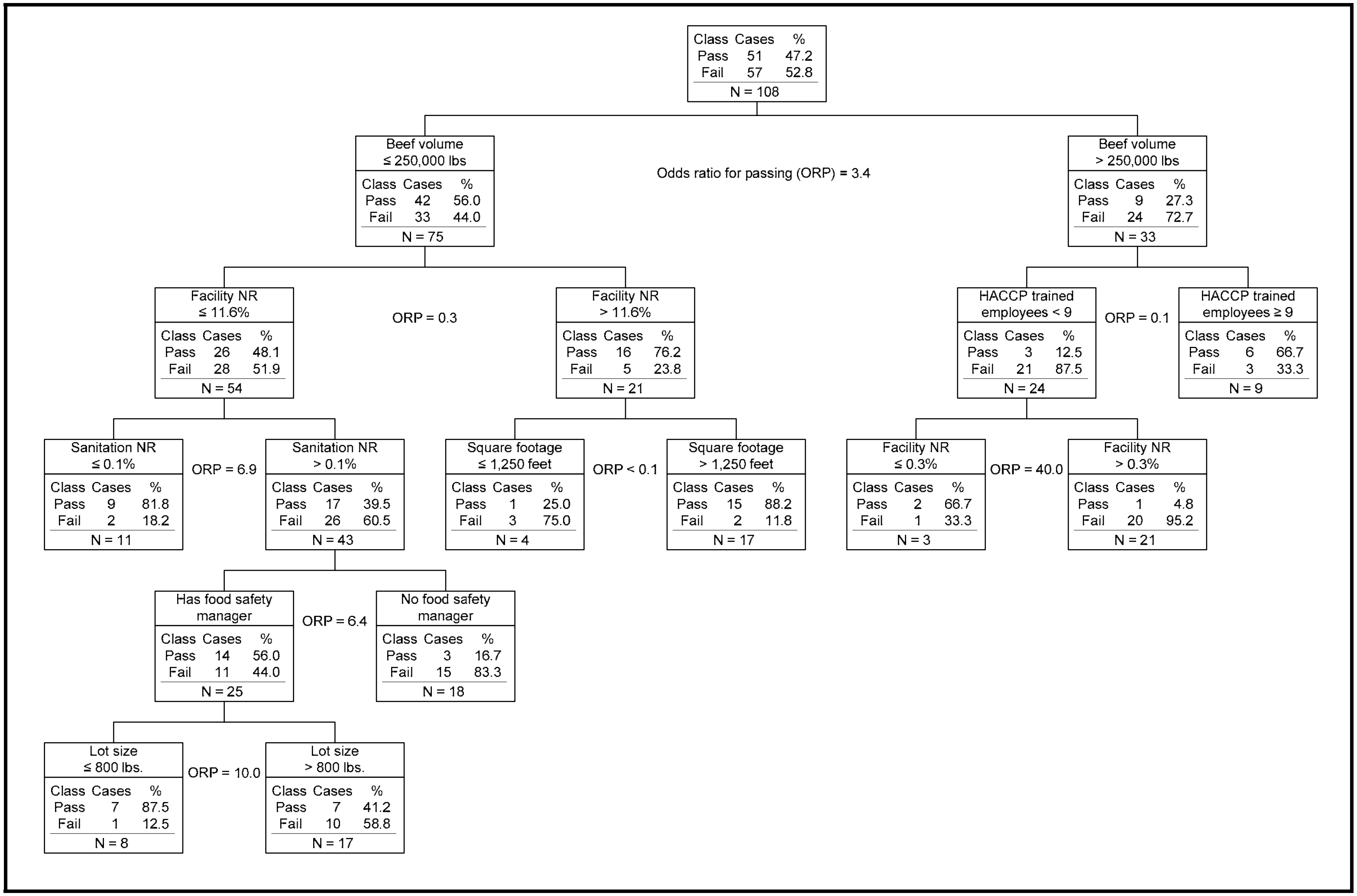
Additional analyses were conducted to determine the relative importance of all variables that might have explanatory power related to *Salmonella* test results in ground beef establishments. The top 5 variables include number of HACCP trained employees, square footage of production space, facility NR rates, volume of beef production, and number of employees in the establishment.

Factor analysis was then conducted to identify sets of continuous variables (or "themes") that may be grouped for further analysis due to their high correlation. The resulting themes relate to establishment size measures (e.g., number of employees and square footage of the production space), NR rate measures (sanitation, facility, and HACCP NRs), other establishment characteristics such as number of days of processing each week and percentage of imported meat inputs; and age of the establishment production space. These themes were further investigated in a logistic regression, but due to the small number of observations and large variability of many of the variables in the model, none of the themes are statistically significant predictors of *Salmonella* test results at the 10 percent significance level.

The final analysis was a stepwise regression procedure in which all continuous and binary variables were included. The results of the stepwise regression indicate the following: establishments that have a specific routine frequency for sanitizing hand or gloves that contact raw meat and poultry are 3.4 times more likely to pass; establishments that use a bioluminescent testing system for preoperative sanitation checks are 4.1 times more likely to pass; establishments that test samples from product contact surfaces, other equipment surfaces, or facility surfaces are less than one-third as likely to pass. Other variables identified in the stepwise regression procedure include two variables that are the same or similar to variables identified in the segmentation analysis: the volume of beef products produced, and whether the establishment provides formal food safety course for newly hired production employees.

In summary, the results of analysis for ground beef establishments suggest the following variables as potential indicators of food safety performance:

- total volume of beef production,
- facility NR rates,
- sanitation NR rates,
- size of the establishment in terms of square footage,
- number of food safety or HACCP trained employees,
- whether the establishment has a dedicated food safety manager,
- the size of production lots produced in the establishment,
- whether the establishment has a specific routine frequency for sanitizing hands and gloves, and
- the types of voluntary testing of surfaces and equipment conducted by establishments.

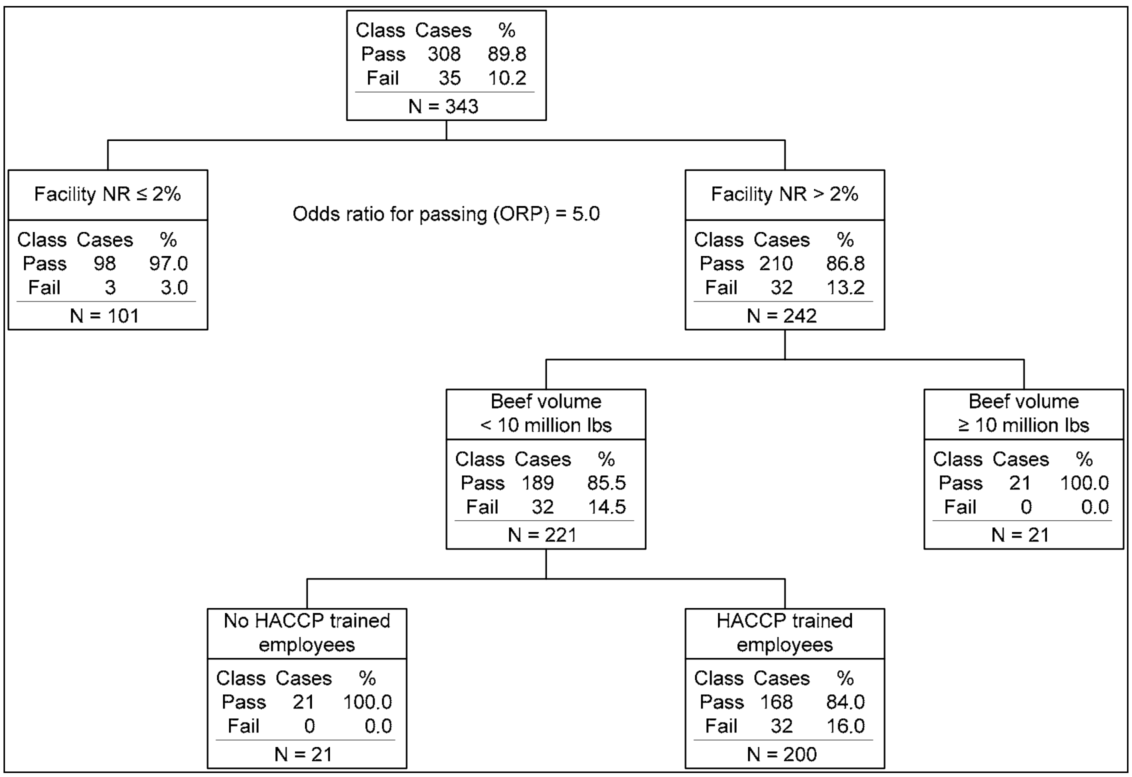


Note: Fail means one or more positive *Salmonella* test results from 2004 through 2006.

Figure E-13. Results of Segmentation Analysis for Establishments that Produce Ground Beef (Including Odds Ratios)

Results of Analysis for RTE Establishments

Figure E-14 shows the results of the segmentation analysis for RTE establishments. Some 60 potential variables were included in the analyses for these establishments. Among these variables, the facility NR rates emerged as the strongest predictor of establishment performance as measured by *Listeria* and *Salmonella* test results. Specifically, among all establishments, the odds of passing (that is, having no positive *Listeria* or *Salmonella* test results from 2004 through 2006) are 5 times higher for establishments with a facility NR rate of less than or equal to 2 percent. Thus, the 343 establishments can be classified into two groups: “lower facility NR rates” on the left side of the tree and “higher facility NR rates” on the right side of the tree.



Note: Fail means one or more positive *Listeria* or *Salmonella* test results from 2004 to 2006.

Figure E-14. Results of Segmentation Analysis for Establishments that Produce RTE Meat and Poultry Products (Including Odds Ratios)

Only three establishments with a facility NR rate below or equal to 2 percent had one or more positive test results; thus, no further analysis of these establishments was conducted. Of the 32 establishments with a facility NR rate greater than 2 percent and having at least one positive test result, all produce less than 10 million pounds of beef products annually, and all have one or more HACCP-trained employees. The result regarding volume of beef products suggests that establishments producing lower volumes of beef products are either producing other products that are more likely to have positive test results, or that these establishments are smaller establishments in general. The result regarding HACCP-trained employees may indicate that the establishments in this group have HACCP-trained employees on staff, but that the training is somewhat less effective compared to other establishments.

Additional analyses were conducted to determine the relative importance of all variables that might have explanatory power related to *Listeria* and *Salmonella* test results in RTE establishments. The top 5 variables include facility NR rates as mentioned above, sanitation NR rates, HACCP NR rates, lot (or batch size), and number of HACCP trained employees. Because relatively few establishments had positive test results over the 3-year period included in the analysis (i.e., only 10.2 percent of the establishments), it was not possible to conduct further statistical analyses to measure the magnitude or statistical significance of the results. However, the results of analysis for RTE establishments suggest the following variables as potential indicators of food safety performance:

- facility NR rates,
- sanitation NR rates,
- HACCP NR rates,
- total volume of beef production,
- number of HACCP trained employees, and
- the size of production lots produced in the establishment.

SENSITIVITY TO PARAMETERS

The previously proposed RCM is comprised of seven parameters: public-health-related NRs; RTE *L. monocytogenes* Alternatives; food safety consumer complaints; food safety recalls; enforcement actions; Salmonella verification categories; and zero-tolerance pathogen test results. Many of those parameters are also proposed to be used in the public health risk-based inspection system discussed in this report. The relative importance of these parameters has been examined, as well as how much weight each factor should be given.

Multivariate analyses are presented here to examine how changing the weight impacts the final RCM.

Analysis of Indicators of a Loss of Process Control

In the above analyses, individual components of the RCM were examined. It is desirable to examine the overall RCM score and how predictive it is of indicators of a loss of process control, as measured by FSIS activities (i.e., NRs, consumer complaints, recalls, enforcement actions, and microbial sampling results). There are some limitations of such analyses, especially due to low supply of available evidence (such as a relatively small number of recorded positive results for *E. coli* O157:H7). Analyses summarized below focus on measuring the utility of RCM scores in predicting a loss of process control as represented by the occurrence of *Salmonella* positives.

Figure E-15 presents AUC scores obtained while predicting an occurrence of a positive result of *Salmonella* test over the next 7 days using scores from RBI algorithms including its component score RCM and Inherent Risk Measure (IRM), as well as combined RBI score (RBIM). The results for seven subcomponents of RCM score are also presented (represented as bar along x-axis). Multiple logistic regression trained on the source data pertaining to NRs and M2K

Salmonella positives was also used. The AUC results of all but logistic regression have been obtained by simply sorting the respective score values across data spanning all establishments and days of analyses and then plotting the ROC curves to reflect output class labels. A perfect AUC score of 1.0 would be obtained by a predictor that would perfectly separate positive from negative cases via sorting. In a more realistic scenario, some of the positive cases will be mixed with negative along the sorted list of records, leading to a lower AUC.

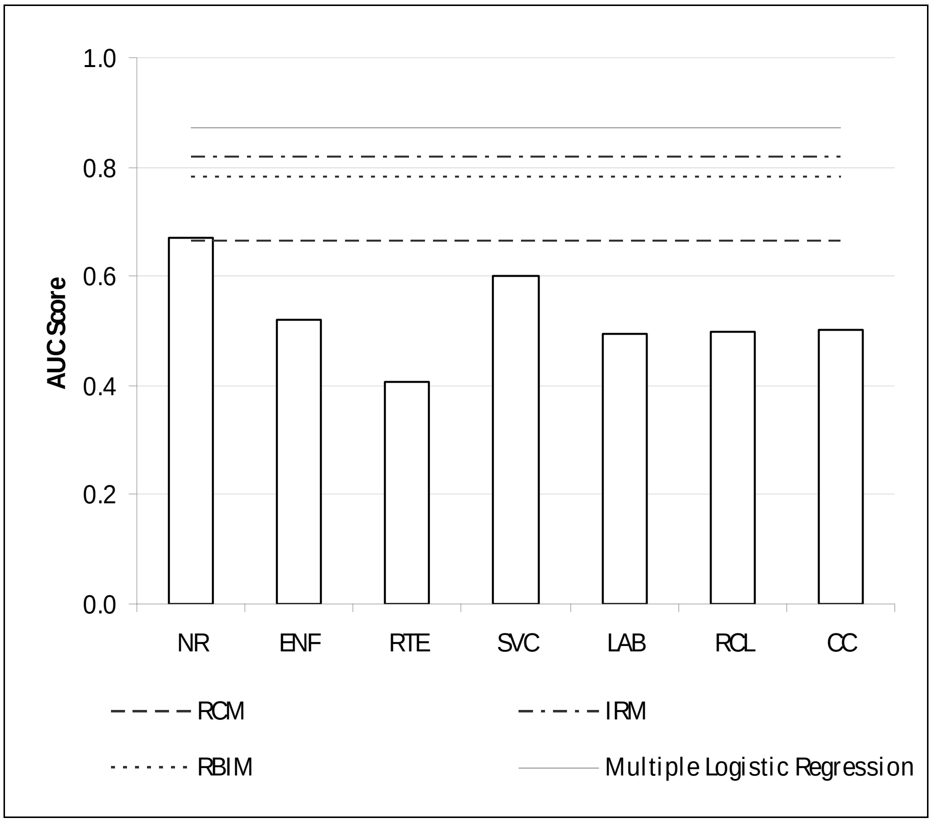


Figure E-15. AUC Scores for RBI Scores, its Component Scores, and From Multiple Logistic Regression

Neither of the individual components of the RCM was found particularly predictive of the occurrence of *Salmonella* positives. The most useful appear to be the scores based on NRs and SVC. The finding that the second of the two scores is somewhat useful in predicting occurrences of *Salmonella* is logical since these measures are specifically designed for the control of this pathogen. An earlier section of this appendix indicated the existence of a useful relationship between NRs, especially specific definitions of NRs relevant to public health, and occurrences of *Salmonella* positives. The AUC of the RTE score is less than 0.5, which suggests that it is negatively correlated with the loss of process control manifested by *Salmonella* positives. That could be explained by the fact that the RTE score focuses on the risks associated with *L. monocytogenes* in RTE products, but it is interesting to note that using an inverse of the RTE score in the formula for RCM might help it better predict occurrences of *Salmonella* positives. After inversion, the expected AUC of the RTE score would be close to 0.6 (i.e., approximately equal to the currently reported AUC for the SVC score). The predictive utility of the combined RCM is similar to that of the NR score, and it is not particularly high. In fact, empirically, IRM based on volume data seems to be more useful in predicting occurrences of *Salmonella* positives than RCM. This is interesting given the fact that the production volume data available for this analysis was limited to one static snapshot of production profile per establishment. Therefore, it could not reflect any changes of production profiles over time, even though such changes would very likely affect the correlations between volume and loss of process control.

Logistic regression is one approach to produce multivariate models of relationships between risk control measures and loss of food safety control. Technically, a trained logistic regression model is a rating classifier which accepts queries composed of multiple continuous input variables and predicts the probability of a given query to be associated with one of the classes of the binary output variable. For example, if the model is trained to predict whether a positive result of a *Salmonella* test will occur next week based on the observation of several parameters of the establishment's past performance (and perhaps its individual characteristics such as size or production profile), it would produce a probability of such an event occurring. The interpretation of that probability measure is essentially analogous to the concept of measuring risk.

In the results presented above, a stepwise logistic regression algorithm was used to illustrate the potential of the multivariate approach. The optimal complexity of the evaluated models was selected using 10-fold cross-validation to ensure robustness against over-fitting, and to establish an objective framework for evaluation of multiple candidate predictive models in the future. In this case, the objective is to identify the components of the smallest subset of variables with the greatest predictive ability (or which minimizes the cross-validation error). The size of that subset would be the optimal complexity.

The training data for this experiment was prepared as follows. Each record corresponded to an individual test for *Salmonella* (as stored in M2K database). It was labeled with the establishment identifier, date, and the outcome (positive or negative) of the test. The outcome was used as the target of prediction. Each record was complemented with a set of input features derived from the M2K and PBIS data. These features included the number of positive results of previous *Salmonella* tests, number of previously conducted *Salmonella* tests, number of all NR citations, number of NRs matching the Industry Coalition definition, and number of NRs of Type 3. Each feature was recorded over 7, 14, 28, 56, 84, and 168 days into the past. Altogether, there were 30 thusly-derived features under consideration by the algorithm. A stepwise logistic regression algorithm was then executed, and the optimal complexity of the resulting model was established via 10-fold cross-validation. The optimal model selected included 13 of 30 available features, the top of which were, subsequently, number of positive results of *Salmonella* tests over the past 168 days, the number of noncompliances defined by Industry Coalition as relevant to public health over the past 168 days, number of *Salmonella* positives over the past 28 days, and number of *Salmonella* tests conducted over the past 14 days.

It is interesting that the model did not select the Type 3 NRs as one of the top features. This can probably be attributed to the high overlap between these NRs and the Industry Coalition grouping. Similarly, production volume was not selected as a top feature. In this case it is probably due to the static nature of the data.

The AUC scores of logistic regression results shown in Figure E-15 outperform each of the RCM component scores and the combined RCM by a wide margin. It also outperforms IRM and RBI; however, the IRM (and therefore RBI) takes into account production volume information which was not considered by this particular logistic regression model. It is likely that the performance of the multivariate approach may be further improved either by using additional informative features (such as production volume or other establishment characteristics) or by employing model optimization methods (such as exhaustive search for the best logistic regression model of a given complexity). Nonetheless, current results already clearly indicate the potential utility of data-driven multivariate predictive modeling in reliable estimation of the expected loss of food safety control.

SUMMARY OF ANALYSES

In this appendix, the presence of positive pathogen results within an establishment has been used as a proxy for measuring loss of process control (and therefore the risk associated with an establishment). The positive pathogen results for *Salmonella* are far more numerous than those for other pathogens and have, therefore, provided a much more robust statistical measure. The weaker results for other pathogens are probably due to the sparseness of the data, especially positive results.

The initial sets of analyses described in this appendix were univariate and were designed to determine the appropriateness of various factors for inclusion in a public health risk-based inspection algorithm. The analyses show that of the tested factors, NRs are the strongest predictor of future process control problems. Properly choosing the subset of NRs to include (excluding the noncleanliness related items) and properly choosing the outcome and evidence window sizes greatly improves their predictive ability. Other factors cannot be shown to be as strong in predicting problems; however, they could be combined into a composite “control measure” component within the algorithm. Further collection of data will improve these analyses.

The multivariate regression tests show that properly choosing a subset of NRs and combining them with the SVC data provides an excellent predictor of process control as measured by *Salmonella* results. The multivariate regression can also be used to determine the best weighting to assign to each factor. The sparseness of data for other pathogens does not a full determination of the ability of these factors to predict other problems. Further data collection will enable this process to be refined.

ATTACHMENT 1: OVERVIEW OF ANALYTIC METHODOLOGY

Lift Statistic: A Measure of Predictive Utility of Parameters

We might know from past experience that if we run a test or a sequence of tests for a specific pathogen at a randomly selected establishment during a given week, there is on average a 2 percent chance (a 0.02 probability) that (at least one of) the test(s) will turn out positive. We would like to know whether there exist some measurable establishment-specific factors which might affect that estimate. If we found these factors in the available data, we should be able to construct data-driven models which should be able to predict the probability of an occurrence of a positive result of the specific pathogen test over a specific period of time in the nearest future at a specific establishment. Such data-driven models could then be used to enable proactive actions by inspectors, and thereby improve public health.

The lift statistic measures the utility of such factors in determining the chance of a positive test result. For example, if we knew that when there was an NR registered at an establishment last week the chance that a subsequently executed *Salmonella* test would be positive was on average 4 times as high as it would be if we did not know whether there was an NR recorded, the lift would be 4. Clearly, it would be useful to know whether there was or was not an NR at an establishment last week, if their occurrence was so highly predictive of the risk of *Salmonella* positives. Any factor that produces a lift significantly above 1.0 is one that should be monitored closely as it frequently precedes pathogen problems (positive results).

In terms of equations, if P(positive test) is the probability of a positive test in general, and

P(positive test | NR last week) is the probability of a positive test given that there was a NR occurrence last week, then the value of the lift statistic from knowing there was an NR is:

Lift(positive test given NR last week)
= P(positive test | NR last week) / P(positive test)

In the example above this might be
= 0.08 / 0.02
= 4

Therefore, lift can be interpreted as an estimate of the increase of risk of certain outcomes of interest (in our example: positive results of microbial tests) given the occurrence of specific facts observed in the available data (in our example: occurrences of NRs).

The probabilities used in the formula above can be estimated from the available PBIS and M2K historical data, by sweeping through all the relevant establishments and through the relevant dates of analysis. One such data extraction cycle is depicted in Figure E-16. For the given establishment and the given day (labeled "today") we look a certain number of days toward the past and check whether there have been issued any specifically defined NRs at the considered establishment within that period of time. We also look a certain number of days ahead toward the future and check whether there were any pathogen tests (e.g., *Salmonella*) conducted and if any of them turned out positive. The lengths of the "looking back" or evidence time window as well as the length of the "looking forward" or outcome window are selectable parameters of the method (note that in the experiments reported above multiplies of 7 days have been used as the widths of these windows in order to discount the day-of-the-week effects on the results). For each such setup we consider what we see a "True Positive" if we indeed do see the sought after NR inside the evidence window and then we also see the positive result of a *Salmonella* test within the outcome window. Please note that the presented method can be used in any context similar to NR vs. *Salmonella* positives which is used here as an example.

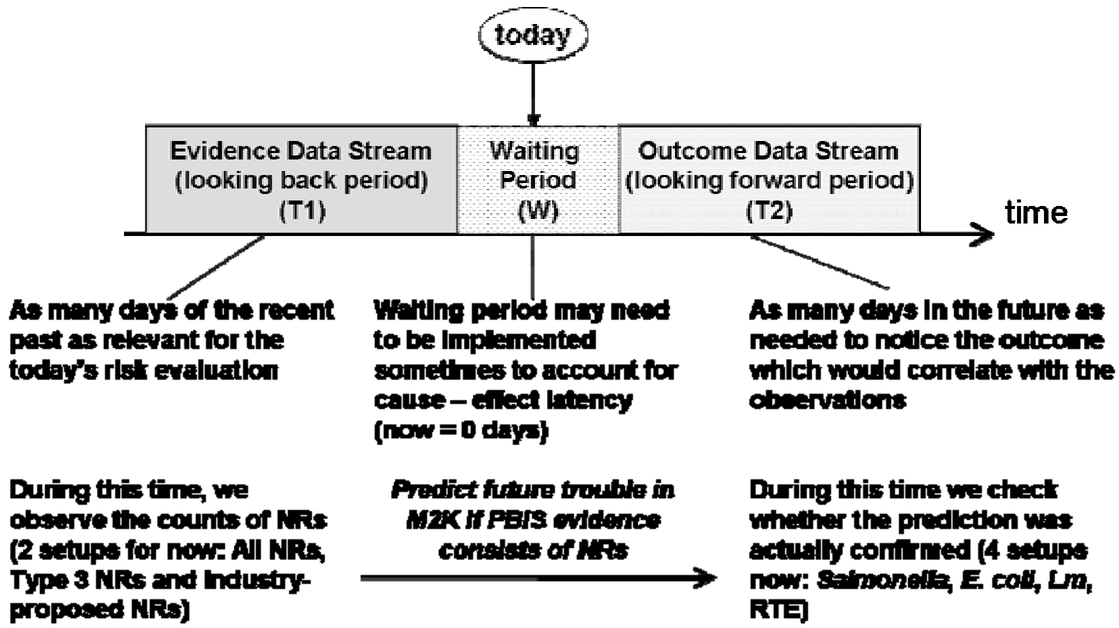


FIGURE E-16 Data extraction cycle.

In Figure E-17, the rows of the main table correspond to the individual establishments and the columns to the subsequent days of analysis. Each cell indicates whether for the given day at the given establishment we have observed an NR inside of the evidence window immediately preceding that day (the result, either "1" – indicating "yes" or "0" – indicating "no" is represented by the first number in the brackets), and whether we have observed a positive salmonella test result over the outcome window immediately following that day (if so, "1" will be the second of the numbers in the brackets). A sequence (0, 1) would indicate a false negative outcome, (1, 1) a true positive, and so forth. The outcomes are then marginalized (aggregated) into contingency tables. A contingency table of binary outcomes and observations is a 2-by-2 matrix with cells storing the counts of the four types of outcomes, respectively true positive, false positive, false negative and true negative. One can imagine creating an aggregate contingency table for individual establishment by accumulating the outcomes over all dates of analysis (these marginal contingency tables are depicted in the dark shading in Figure E-17), or the aggregation can be performed on a day-by-day basis (for each day across all establishments, depicted in the patterned shading in the figure), or it can be done globally (across all establishments and all days). The last option (global) is the one of chosen for the purpose of the tests reported in this appendix.

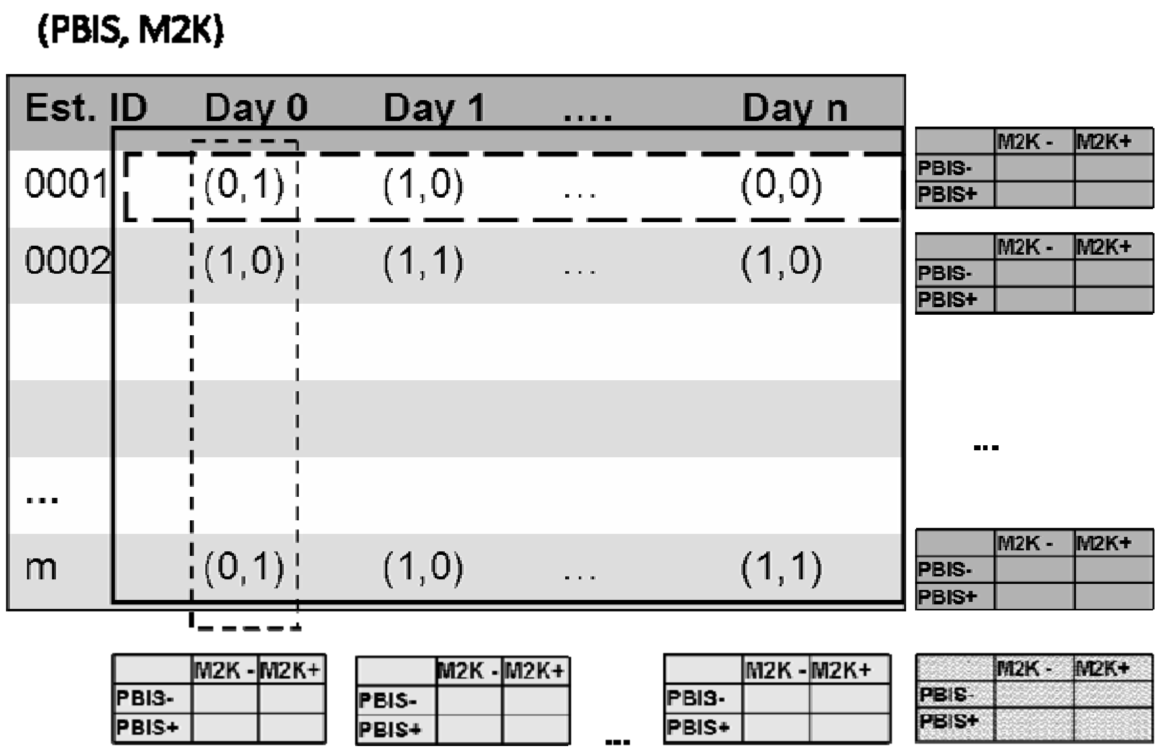


FIGURE E-17 Joint contingency table to detect M2K result upon PBIS occurrences in terms of 'lift'.

Once the joint contingency table is assembled, the probabilities needed for lift estimation can be derived directly from the aggregated counts as follows:

P(Positive *Salmonella* test in the near future | NR in the recent past) = TP / (TP + FP)

P(Positive *Salmonella* test in the near future) = (TP + FN) / (TP + FN + FP + TN)

Here, TP = count of true positive cases recorded in the aggregate contingency table, FP = count of false positive cases, TN = count of true negative cases, and, FN = count of false negative cases. Then, as shown before, the equation for lift is:

Lift = P(Positive *Salmonella* test in the near future | NR in the recent past) /
P(Positive *Salmonella* test in the near future)

Intuitively, the lift statistic measures a relative benefit of paying attention to occurrences of NRs in predicting occurrences of *Salmonella* positives, versus ignoring the information about the NRs in doing so. A lift value of 1.0 indicates no benefit. Values greater than 1.0 suggest a potential utility in using NRs to predict positive *Salmonella* tests. Values of lift smaller than 1.0 would suggest that the presence of NRs is negatively correlated with the presence of positive test results in the immediate future.

The analyses presented in this appendix make use of the lift statistic mainly to check whether there is evidence of correlational dependencies of observables (such as occurrence of NRs of certain types over the recent past) and the outcomes indicating a potential risk to the public health (such as the positive outcomes of microbial tests). High and statistically significant values of lift suggest a potential utility of the specific observables in estimating risk, although they do not necessarily indicate causal relationships between the observables and the outcomes. It is important to mention that the lift statistic as defined above focuses mostly on the positive outcomes of tests. In order to measure the overall performance of any predictor it is necessary to also consider the impact of negative cases on the accuracy of prediction. A convenient way of accomplishing that is to construct ROC (Receiver Operating Characteristic) graphs and compute AUC (Area Under the Characteristic) scores which quantify the ability of a predictor to accurately discriminate positive from negative outcomes based on the available observations

The analyses for each of the discussed pairs of data streams in this appendix have been performed for each of 25 combinations of evidence and outcome window widths selected from the following list of choices: 7, 14, 28, 56 and 84 days. Where enough data was available and the lift appeared significant, both ROC and AUC were computed. Unless otherwise noted only statistically significant findings are reported.

Testing Significance of the Lift Statistic and AUC Scores

The analyses discussed in this appendix produce aggregate contingency tables for a number of combinations of the evidence window sizes and the outcome window sizes. From each of these aggregate contingency tables, true positive rate, false positive rate, and lift can be easily computed. By holding the evidence window fixed and sweeping through different outcome window sizes (or vice versa) one can obtain a ROC curve and compute its AUC score. It is entirely possible that the lifts and AUC scores so obtained may be due to pure chance and they may not differ substantially from the results which could be obtained if the data was random. In such a case, any supposed evidence of a correlational relationship between NRs and *Salmonella* positives would have to be dismissed. Randomization tests of significance are therefore conducted in order to verify the original set of results against their deterministicity.

One approach to testing whether the particular values of lift or AUC have been obtained by chance is to randomize data in a way that would break the supposedly existing relationship between the observables (e.g., PBIS data) and monitored outcomes (e.g., M2K microbial test results) and then to re-compute the values of lift and AUC. If the re-computed values would not be substantially and systematically different from those obtained originally, one would not consider the original results trustworthy.

In the NR vs. *Salmonella* example, we first randomly shuffle the positive labels of the *Salmonella* test results among all of the tests that were performed (across all considered establishments and dates), so that some tests labeled as negative in the original data will turn positive and vice versa. Note that in this test the test dates and the total number of tests as well as the total number of positive results remain intact. Then, from the randomized data we extract the aggregate contingency table and compute lift and AUC in the exactly same way as it is done for the original undisturbed data. The lift and AUC so computed might be higher (better) or lower (worse) than the results obtained for the original distribution of positive tests. If we perform this shuffling-and-computing many (say 999) times, we will have lift and AUC values for 1,000 distributions of positive test results: the one set from the original distribution and the others from the 999 randomly generated distributions. We can count how many of these distributions have results better than or equal to the original lift or AUC value, respectively. (The count will be at least 1, since we include the set of results obtained for the undisturbed data to the pool.) The fraction (count /1000) becomes then an estimate of the probability of observing a result at least as good as that computed from the original distribution just by chance. If this probability (a p-value) is very low (say, less than 0.05), we would have some confidence in that the observed distribution is actually not due to random chance, and that there is in fact a non-accidental relationship between occurrences of PBIS NRs and an increased probability of a subsequent M2K positive test. A second (less conservative) test can then also be performed in which the pathogen test dates are also varied.

Note that the confidence intervals can be asymmetrical since we do not make any assumption about the shape of the randomization distribution. The intervals are calculated nonparametrically. Given a sample of randomized scores, we pick the top 2.5 percent and the bottom 2.5 percent and we obtain the confidence limits thusly. It sometimes occurs that among these synthetic scores 2.5 percent or more correspond to zero lift. Then the lower confidence limit ends up being set to zero (lift cannot be negative).

Some particularities of the analytic results obtained through lift and ROC analysis might be due to the non-random selection of establishments under consideration. In order to measure the sensitivity of the lift and AUC results against random fluctuations of the composition of the set of considered establishments, we execute the following bootstrap procedure. For each establishment, we construct its contingency table by counting the co-occurrences of NRs and *Salmonella* test results in their respective time windows, over the time span of the considered data. Then, a large number of times (say S-1=999 since we add the original set of results to make the total number of samples S=1000) we repeat the following: randomly sample (with replacement) N establishments (here N is the total number of establishments under consideration) and aggregate their individual contingency tables into one table from which we then compute lift and AUC values. Note that each of those S-1 random samples of N establishments may include repetitions of some establishments whereas some others may not be represented at all. If the performance of the original set of establishments was not internally consistent in a way that could be reflected through their contingency tables, we would see a wide variability of the lift and AUC scores obtained via such randomization process. Otherwise the variability obtained would be small. After collecting the S results we report the values of the resulting statistics (lift and AUC) corresponding to the mean between the K^{th} and $(K+1)^{th}$ highest scores as the upper (1-2K/S)*100 percent randomization confidence interval limit (K=25 for 95 percent intervals), and the mean of the K^{th} and $(K+1)^{th}$ lowest scores as the lower randomization confidence interval limit.

Overview Of Data Sources

M2K is a USDA system that contains the results of pathogen tests performed on samples taken at establishments. It contains data from January 2005 to the present. For these analyses we used a set of this data that spanned January 2005 through March 2007. Table E-18 summarizes the number of data points for each pathogen by project code and also the total number of results (positive and negative). The column heading is the source of the data categorized by project code. The row title on the left hand side is the analysis category used in the lift calculations.

Table E-18 Summary of Pathogen Test Results in M2K from January 2005 Through March 2007

	Project								**Total**	
	Salmonella		***Lm***		***E. coli***		**RTE**			
Analysis	**Neg.**	**Pos.**	**Neg.**	**Pos.**	**Neg.**	**Pos.**	**Neg.**	**Pos.**	**Neg.**	**Pos.**
Salmonella	96,291	5,642	0	0	1,743	0	30,069	12	128,103	5,654
Lm	0	0	3,549	5	0	0	33,423	288	36,972	293
E. coli		0	0	0	28,556	53	1,433	0	29,989	53
RTE	0	0	0	0	0	0	64,925	300	64,925	300

The following are the project codes that were used in the analysis:
Salmonella: HC01
Ecoli: MM45, MM45R, MT03, MT04, MM45F, MT50, MT52

LM: RLMCONT, RLMPROD
RTE: ALLRTE, INTCONT, INTPROD, RTE001, RTERISK1

PBIS is a USDA system that contains results of inspections performed at establishments. The system has undergone several refinements and changes since its inception and therefore it is not possible to utilize all of the data within PBIS in a single analysis. Clean, stable data used for these analyses from within PBIS begins in January of 2006. For this reason factors that require analysis of the combined M2K and PBIS data can only be performed on the subset between January 2006 and March 2007. Table E-19 summarizes the number of establishments that are present in the intersection of these data sources for different groups of NRs (within PBIS) and pathogen tests (within M2K).

Table E-19 Summary of Number of Unique Establishments that Are Present in the Intersection of M2K Data and PBIS Noncompliance Data from January 2006 Through March 2007

Type of NR	*Salmonella*	*E. coli*	*Lm*	RTE
All	3,382	1,823	2,349	2,349
Industry-proposed	3,159	1,715	2,170	2,170
Type 3	3,194	1,715	2,217	2,217

The recall data used in these analyses spanned the time from March 2004 to March 2007. All recall data are extracted from FSIS recall website located at http://www.fsis.usda.gov/ Fsis_Recalls/. Table E-20 summarizes cleaned recall data by reason.

Table E-20 Summary of Recall Data by Recall Reason from March 2004 to March 2007

Reason for Recall	Number of Recalls			
	Class 1	**Class 2**	**Class 3**	**Total**
Foreign material	7	3	1	11
E. coli contamination	20	0	0	20
Lm contamination	49	0	0	49
Pathogen contamination	1	0	0	1
Misbranded	3	0	4	7
Mislabeled	14	3	2	19
Pesticide contamination	0	1	0	1
Adulterated	1	0	0	1
Salmonella contamination	3	0	0	3
Bug contamination	2	0	0	2
Allergen	7	5	0	12
Undercooked	6	0	0	6
Total	113	12	7	132

The CCMS data available spanned the time from April 2006 to September 2006. Table E-21 summarizes the data in the OPEER and EPI cuts of these events.

Table E-21 Summary of CCMS Data from April 2006 to September 2006

Measure	OPEER Cut	EPI Cut
No. of instances in raw data	423	47
Less: No. of instances discarded as not enough establishment identification information available	140	3
No. of instances ended up in analysis	283	44
No. of unique establishments	163	35

A record of enforcement actions by establishment is also kept at USDA. This data contains 59 NOIEs issued to 58 unique establishments during the period from April 2006 through September 2006. This data is collected according to the date of the notice and is stored in a table in the data warehouse.

References

Cates, S.C., S.A. Karns, J.L. Taylor, C.L. Viator, and P.H. Siegel. April 2006. "Survey of Meat and Poultry Processing-Only Establishments." Report prepared by RTI International for the U.S. Department of Agriculture, Food Safety and Inspection Service, Washington, DC. Available at http://www.fsis.usda.gov/PDF/SRM_Survey_Meat_&_Poultry_Processing_Only_Plants.pdf