RESPONSE TO PEER REVIEW OF THE *E. COLI* O157:H7 RISK-BASED SAMPLING ALGORITHM

(May 2007)

In August 2007 the *E. coli* O157:H7 risk based sampling algorithm was formally peer reviewed by experts from outside the federal government. An outside contractor who chose five reviewers based on the expertise required to best evaluate the algorithm orchestrated the peer review. The panel included professionals from academia, industry, and foreign government with expertise in quantitative risk assessment, public health statistics, and statistical design of sampling programs (see Appendix A). The reviewers were asked to respond to six specific charges that target the key elements of the algorithm (see Appendix B). All additional comments/input were also considered. This document summarizes the comments made in response to each of the charges and the corresponding answers (where relevant) from the Risk Assessment Division (see Appendix C for the complete reviews).

Table of Contents

Overview

The reviewers were in favor of the Food Safety Inspection Services (FSIS) implementing the *E. coli* O157:H7 risk based sampling algorithm. All five reviewers agreed that the algorithm is based on a scientifically sound approach and represents a positive step for FSIS. Reviewers agreed that the probabilistic (Monte Carlo) structure of the algorithm is an effective framework for allocating samples based on risk and that the algorithm will help FSIS allocate sampling resources to the "best possible advantage". All reviewers stated that the algorithm will be greatly improved when it is expanded to include establishment interventions and testing programs—scheduled for April 2008. Reviewers also commented that FSIS needs to better document and justify both the choice of sampling probability weights and the use for a multiplicative approach for combining the individual weights. In addition, there were questions about the approach used to weight volume vs. hazard and there interdependence. All of these comments have been addressed through changes to the algorithm itself and improvements to the documentation. Detailed responses to all comments as well as the full-length reviews and reviewer biographies are contained in this report.

Response to Comments

CHARGE 1

Is the overall approach for allocating samples through a probabilistic draw sound? (a random draw where the probability of selection is weighted by risk) If not, what problems exist and how should they be addressed?

Charge 1 Commendations

Reviewer 1: ***"****I find the approach of using a random draw with weighted probabilities for selection of plant to be sampled is an appropriate method to meet the objectives of this program and the Agency. This addresses the issue of using Agency resources to best possible advantage to find sources of E. coli O157:H7 in ground beef and trim… The Agency should be commended for taking this approach in sampling ground beef and trim in effort to reduce the incidence of E. coli O157:H7 in the nation's food supply as well as making better use of their resources to monitor food safety*"

Reviewer 2: "*The general approach is sound overall. This kind of sampling is quite common in both scientific and regulatory practice, and so most users of the method*

should find it understandable. Embedding the approach inside a Monte Carlo routine is the correct thing to do, and the authors have correctly executed (in the code) the Monte Carlo procedure… The overall approach is a form of Analytic Hierarchy Process, which also has a long history of use in decision-making."

Reviewer 3: "*The approach employed in the developed algorithm is statistically sound and is the right way to improve over the existing practice. If implemented correctly, this will be a cost-effective and more accurate alternative to the current practice in controlling and monitoring the E. coli O157 in the beef industry.***"**

Reviewer 4: "*The approach for allocating samples through a probabilistic draw is sound and has a great potential to improve the efficiency of the surveillance program to protect consumers from E. coli O157:H7 infection.*"

Reviewer 5: ***"Yes, the approach is sound and well justified based on known risk factors."***

Charge 1 Critiques and Responses

Reviewer 1: NA

Reviewer 2: *"The only quibble I would have at this very general level is their choice of a multiplicative function for the 2 attributes they consider (volume and past results). There is nothing wrong with that algorithm per se, and it does have the good feature of giving each of the two attributes roughly equal weight in the sampling. But there are other algorithms they might have used, such as specifying a measure of "importance" to each of the two attributes and then taking a weighted sum (score from attribute 1 times the importance of attribute 1 + score from attribute 2 times the importance of attribute 2 + …)… I am not at all suggesting they change the algorithm, but it would be helpful to the reader to understand why they chose this particular multiplicative function..."*

Response: We agree with the reviewer that there are many valid choices available for combining the algorithm attributes. The multiplicative function is justified in this case by

Risk = Exposure*Hazard

Or, in the case of the risk based sampling algorithm,

Risk = (production volume of ground beef)*(likelihood of O157 contamination)

Reviewer 3: NA

Reviewer 4: NA

Reviewer 5: *"A system needs to be implemented to ensure that the algorithm used in a specific year incorporates the most recent/relevant information. I recommend that FSIS develops a plan and a data analysis strategy to do this to minimize possible criticism that the sampling probabilities are outdated."*

Response: The sampling results and production volume data for each run of the algorithm will include the 12 months of data collected up to the previous month. FSIS is still planning the survey that will be used to collect information on plant interventions and testing programs. We agree with the reviewer that it will be important to collect this information as often as is feasible given inspection resource constraints. Currently, we are considering updating plant profiles in the algorithm on a quarterly basis. The frequency of these updates will depend on the frequency FSIS is capable of collecting updated plant profile data.

CHARGE 2

Evaluate algorithm source code and mathematics. Are the techniques (mathematics and equations) appropriate? If not, provide alternatives. The reviewer should examine and verify that the data analysis and source code are accurate.

Charge 2 Commendations

Reviewer 1: "*The mathematical approach for this program seems to appropriate for making risk based decisions.*"

Reviewer 2: NA

Reviewer 3: "*Yes. I think the mathematics and equations are correct.*"

Reviewer 4: "*The techniques are appropriate and correctly implemented in the source code…The model is very user-friendly and well documented. The source code is flexible and structured in a way which allows modification of the algorithm also by less experienced visual basic programmers*."

Reviewer 5: NA

Charge 2 Critiques and Responses

Reviewer 1: NA

Reviewer 2: "*Based solely on the document, it is impossible for the reader to understand how the Volume Score is calculated (the same is not true of the Hazard Score). This then makes it impossible for the reader to confirm the results in Table 18. There must be a complete set of calculations for at least a few example facilities (establishments or sites) showing exactly how the probability ranges were determined. Let me suggest that the authors include a much improved flow diagram of the computational steps, showing the*

decision structure…The main problem with the Volume Score lies in the reader determining whether the four criteria shown in the box diagrams are intended to apply to individual facilities and their probability ranges (the ranges used in the Monte Carlo selection process) or only to volume categories."

Response: Due to issues raised here and elsewhere in the peer review, we have revised the volume score calculations. The new method is greatly simplified and addresses the above concern as well as concerns raised elsewhere in the peer review (section 3B, reviewer 2, and section 5 reviewer 4). The revised method is explained in detail in the updated report. Briefly explained, the algorithm now calculates volume score according to a simple linear "scaling down" of the actual volume scale. For example, we estimate that the largest plants produce ~750x more product by weight each day than the smallest plants. Based solely on this data, FSIS would sample the largest producers 750x more than the smallest—and the intermediate producers at proportionate frequencies. However, there are constraints on our sampling program that make this direct relationship of sampling to volume unfeasible. For instance, given the number of samples available for the program, this would mean ~ 1,000 plants would go years without being sampled a single time while the burden on inspectors in large plants as well as the large producers themselves would be unreasonable. The algorithm solves this problem through a "scaling factor" that reduces the 750x difference to a level that risk managers have determined provides a feasible level of sampling for the program. Scaling reduces the actual difference between production categories proportionately according to the relationship

$$\text{Volume Score for Establishment } i = S_L + \left(\left(V_i - V_4\right) \Big/ \left(\frac{\left(V_1 - V_4\right)}{\left(S_H - S_L\right)} \right) \right)$$

V_i = Production volume of establishment i, there are 4 volume categories, 1 produces most and 4 least
V_1 = Production volume of establishments in category 1
V_4 = Production volume of establishments in category 4
S_L = Lowest score of the scale
S_H = Highest score in the scale

Reviewer 3: NA

Reviewer 4: "*In addition to the documentation provided, a list of the variables used in the visual basic code and their meanings would be helpful."*

"Details on suggestions for alternative mathematical solutions for the algorithm are discussed under e."

Response: The VB variables are listed in the source code.

Reviewer 5: "*On page 38, why give a value of 0 if there are no data available? What are the reasons for missing data – no recording of information, stopped production?*

Response: The most common reason for a plant to have no volume data available is that they are a new establishment. The algorithm identifies those plants with no volume data, codes them as "0", and then assumes they have the highest volume score possible until data is available--- usually after the first sample is taken since this data is collected every time a plant is sampled.

If a small establishment is not sampled once in the first 11 months, will its sampling probability in the twelfth month be 1?"

Response: We have added a "minimum sample rule" to the algorithm. The allowable minimum number of samples/year was determined by FSIS risk managers to be 3. The algorithm begins by allocating a single sample to each plant that has not had the minimum number and then allocates the remaining samples according to sampling probability.

"The odds ratio (OR) calculations (Tables 9 and 10) are not exactly correct because a comparison group of all samples is used. Standard epidemiologic practice dictates that the 2 exposure categories should be mutually exclusive, e.g. factor positive versus factor negative if the exposure is binary. However, given the rareness of the outcome, a positive O157 result, they do not change dramatically"

"Also, on pages 20 to 22, I strongly suggest that the seasonal data (esp. data in table 14) be presented as OR to be consistent with presentation of data for sample testing history. Based on my calculations, the OR is 2.15 for a comparison of the two seasonal categories. One final point about the OR calculations -- the confidence intervals do not account for the clustered sampling design and hence, are narrower than they should be."

Response: We have changed the above calculations accordingly and adapted the relevant tables in the report.

CHARGE 3

Does the algorithm accomplish the three objectives?
A) To increase the proportion of FSIS samples taken at establishments that are more likely to produce product contaminated with E. coli O157:H7.
B) To allocate FSIS resources more efficiently by verifying a greater portion of the U.S. trim and ground beef supply with the same number of samples as the current program.
C) To verify all eligible establishments at a reasonable frequency regardless of an establishment's production volume, interventions, or predicted public health risk associated with their product.

Please be specific. If so describe why and if not suggest how it could be altered to better achieve the described objectives.

Charge 3A Commendations

Reviewer 1: "*The algorithm does aid in increasing the proportion of samples taken at establishments that are more likely to produce product contaminated with E. coli O157:H7. This will improve as the other factors like seasonality and establishment practices are included in the model."*

Reviewer 2: "*My answer here is that the methodology does indeed accomplish this. This is evident both in the structure of the methodology and in the results in Figures 6 through 10."*

Reviewer 3: *"Compared with the current system, which assign every establishment same weight, the proposed algorithm indeed increases the proportion of the large establishments to be tested, and this translated into a bigger portion of the total U.S. beef production to be tested more frequently than before"*

Reviewer 4: ***"****All three stated objectives are accomplished by the algorithm."*

Reviewer 5: *"This is achieved by increasing sampling in establishments with a sample history and high production output."*

Charge 3A Critiques and Responses

Reviewer 1: NA

Reviewer 2: NA

Reviewer 3: "*If the algorithm is fully implemented, i.e. after the proper score for the establishment's intervention and testing programs are incorporated to the procedure, it will be a more complete/accurate algorithm. As it currently stands, the weights in the algorithm are only determined by two factors, the production volume and recent test results. There are some important details left unspecified, e.g., how the season and establishment's prevention programs will be weighted in the algorithm and how sensitive/reasonable the new weights will be. Furthermore, the composition of weights may also need evaluated further, e.g., one may consider additive form of the composted weights or the logarithm transformed forms*"

Response: The algorithm will be peer reviewed again once we have incorporated the data on plant interventions and testing programs currently being gathered by FSIS.

Reviewer 4: NA

Reviewer 5: NA

Charge 3B Commendations

Reviewer 1: NA

Reviewer 2: "*The methodology clearly produces a more efficient sampling scheme than that used currently, in the sense of allocating limited sampling resources to target the points of most effective intervention in the food supply."*

Reviewer 3: "*Objective 2 is accomplished"*

Reviewer 4: ***"****All three stated objectives are accomplished by the algorithm."*

Reviewer 5: NA

Charge 3B Critiques and Responses

Reviewer 1: *"… if looking at volume alone, the smaller plants (<1,000 lbs/day) are still being sampled more frequently on a per pound basis than are the larger plants… Using the algorithm to assign probability based on volume alone, the small plant would be sampled an average of 4 times with each sample representing 60,000 lbs of product. The large plant would be sampled 11 times with each sample representing 5.5 million pounds of product. This is going in the right direction of being equitable sampling, but is still a long way off."*

Response: The reviewer is correct. However, there is more to consider in a risk based sampling program than the analyses. As discussed in response to earlier comments (Charge 2, Reviewer 2), resource burden, and feasibility must also be considered. For instance, if we were to adjust sampling frequency directly proportionate to production volume, given the current number of available samples, ~ 1,000 plants would likely go unsampled for years while the largest producers would be sampled at an unreasonable level for both the FSIS inspectors and the producers themselves. Therefore, we have created an algorithm that allows risk managers to "scale down" the actual volume scale to one FSIS can reasonably utilize (see response to Charge 2, Reviewer 2 for details).

Reviewer 2: *"I am not, however, fully convinced that the first criterion in bounding the Volume Score, the one related to fraction of establishments and fraction of the total supply, is needed… This objection might be overcome by adding a paragraph describing the implications of this criterion on the sampling, perhaps showing an example of the sites selected when the criterion is not employed and when it is added in*."

Response: Due to issues raised here and elsewhere in the peer review we have revised the volume score calculations (see response to Charge 2, Reviewer 2 for details).

Reviewer 3: NA

Reviewer 4: *"For objective two (verifying a greater portion of the supply), it is difficult to evaluate how much of the supply can be verified with the program. One sample per establishment will be less representative of the total daily production of a large establishment compared to establishments with a lower production."*

Response: We agree with the reviewer. Although FSIS can estimate daily production volumes for most plants, it is difficult to determine what proportion of an establishment's product is "verified" to be free of O157:H7 when a single sample is taken from a single production lot. That being said, verifying more lots from plants that produce more of the total supply means FSIS has increased confidence in the *E. coli* O157:H7 control in more of the US ground beef supply.

Reviewer 5: ***"****This assumes that the risk-based estimates are unbiased and not confounded by other unknown factors. The greater the magnitude of the risk factor(s), the more beneficial the use of a risk-based approach to sampling."*

Response: We agree. In order to guard as much as possible from the confounding effects of unknown and/or unaccounted for risk factors we have designed the algorithm to be probabilistic rather than deterministic. All plants are still sampled with some probability and so their controls are verified to some degree whether we have identified high-risk practices in the establishment or not.

Charge 3C Commendations

Reviewer 1: "*It meets the objective of allocating Agency resources more efficiently and verifies all eligible establishments at a reasonable frequency."*

Reviewer 2: "*This part of the approach meets the goal entirely through a bounding or constraint that requires each site to be monitored at least once in some prescribed period."*

Reviewer 3: "*Objective 3 is accomplished in both newly proposed and existing algorithms."*

Reviewer 4: ***"****All three stated objectives are accomplished by the algorithm."*

Reviewer 5: ***"****This is achieved by ensuring that all establishments are sampled."*

Charge 3C critiques and responses

Reviewer 1: NA

Reviewer 2: "*This was, however, one place where I could not fully understand how the code was accomplishing this part of the approach. That needs to be explained more fully in the document*."

Response: The algorithm has now been programmed with a maximum and minimum number of annual samples such that no plant can receive more than 2/month and no plant can receive fewer than 3/year.

Reviewer 3: NA

Reviewer 4: ***"****Objective three is relatively subjective, because there is no clear definition of what is meant by a reasonable frequency. In the suggested program, a median of 4 samples per year are collected from establishments in the lowest risk category. While this should be sufficient to ensure a reasonable frequency, a better definition of this objective would help to improve the algorithm. For example it could be stated that at least X% of the samples should be allocated randomly, at least n samples should be available for establishments in the lowest risk category, or the expected number of samplings should be at least X per year for all establishments."*

Response: FSIS risk managers have determined that given current resources available a reasonable minimum number of annual samples is 3. The algorithm has now been programmed so that no plant will have fewer than 3 samples per year.

Reviewer 5: *"Reasonable frequency" is not defined but might be as low as a single sample in some establishments. The changes in the program sampling are clearly shown in Figures 6 to 10."*

See response to Reviewer 4 above.

CHARGE 4

The algorithm uses four general areas to assign the overall risk of causing O157:H7 illness (production volume, sample history, season, and establishment practices). Comment specifically on the use of each of these areas. Are there additional factors that should be considered?

Charge 4 Commendations

Reviewer 1: *"The proposed method of weighting the probability for selection currently relies on volume and sample history, but as more information is gathered on establishment practices that will effectively reduce the incidence of E. coli O157:H7 on beef, I feel that this algorithm could be a very powerful tool in focusing the testing efforts of the Agency."*

Reviewer 2: *"…the exposure measure is quantified entirely by the production volume. Given that one cannot say much about where the product will end up, I think this is probably the most precise measure one can use. And so I fully support the idea of using it to quantify v in the probabilistic sampling calculations… The authors then intend for three factors to go into the measure of hazard: (1) past sampling results, which provide an indication as to whether a particular facility has had a problem with contamination in the past; (2) seasonality, since there is good evidence that occurrence of contamination is dependent on season; and (3) introduction of practices that might reasonably be expected to reduce the occurrence of contamination (or whose absence will increase the occurrence). These seem to me three very reasonable factors to include."*

Reviewer 3: *"The selection of the four areas to determine the relative weights in the proposed algorithm is appropriate."*

Reviewer 4: NA

Reviewer 5: ***"****First, the use of production volume is well justified since large establishments produce the greatest volume of servings and theoretically, more contaminated servings given a constant individual risk of a contaminated serving. Second, the choice of sample history was well justified based on the odds ratio estimate of 4.86. However as the authors indicate on page 18, collection of additional follow-up samples over time will greatly improve the ability to draw conclusions about the importance of sample history E. coli O157:H7 risk. Third, the seasonal risk of human illness associated with consumption of ground beef products closely parallels the seasonal risk of E. coli O157:H7 positive samples and hence, there is sound justification for inclusion of season in the sampling design."*

Charge 4 Critiques and Responses

Reviewer 1: "*As the other risk factors of history, seasonality and establishment practices are included, the algorithm will give a better recommendation in sampling based on the actual probability of finding E. coli O157:H7…Another risk factor to include might be geographic location of the plant or source of the cattle."*

Response: We agree with the reviewer. Inclusion of establishment practices in the algorithm has the potential to improve its utility. FSIS has recently finished the first round of a survey (Dec 2007) that collects the needed data from beef producers and we expect to complete our analysis and implementation in the coming months. As mentioned elsewhere in this document, the revised algorithm will be submitted to further peer review. At this time, FSIS will not be allocating samples according to the increased O157 occurrence in the warmer seasons. This is a laboratory resource issue.

Reviewer 2: "*What I am not at all clear on is how these additional two factors (seasonality and plant practices) will go into the complete algorithm. I believe they all fall under the Hazard Score, at least in principle, which suggests the three factors will be collapsed into one Hazard Score, to then be combined with the Volume Score in precisely the manner already executed. . But the authors then face the difficult decision as to the manner, computationally, for folding these three factors together*."

Response: Yes, the additional factors will be added to the hazard score. Please see the above response to Reviewer 1 for more details.

Reviewer 3: NA

Reviewer 4: "*There is a very high correlation between production volume category and plant probability points (Spearman's rho for January 2006 data = 0.997). The current algorithm results in a random sample weighted by production volume, with re-testing of the few establishments with a history of contamination. This result could have been achieved with a much less sophisticated model. In order to utilize the potential of risk-based sampling, additional risk factors need to be taken into account. Of the factors discussed, establishment practices such as process control, interventions, employee training and testing programs are the most important ones to include. This helps to identify establishments with a decreased risk of contamination. In addition, recording these factors will provide an incentive to establishments to improve their manufacturing practice. If it is practical to vary sampling intensity by season, this risk factor should also be included as described in the report. Participation in quality assurance programs and geographic location of the establishment or the suppliers could be additional risk factors to consider."*

Response: We agree with the reviewer. Please see our response above to reviewer 1.

Reviewer 5: ***"****Finally, establishment practices that reduce the prevalence of O157 will be included. Since these are not available for evaluation, then I can only indicate that philosophically this is a good choice."*

Response: Please see the response above to reviewer 1.

CHARGE 5

Each of the four areas (production volume, sample history, season, and establishment practices) is assigned a weight in the algorithm to determine the overall probability of sampling for each establishment. In some cases weighting is proportionate (i.e. seasonality and sample history are weighted proportionate to the increased probability of detecting contaminated lots) while in other cases the weighting needs to be balanced with the overall objectives of the program (i.e. production volume weighting is more complex due to additional sampling concerns). Comment specifically on the method of weighting each area. Is the weighting consistent with the three objectives?

Charge 5 Commendations

Reviewer 1: "*The weighting of the volume is balanced with the overall objectives of the program. This is consistent with the objectives of the program to verify all eligible establishments at a reasonable frequency. The weighting of sample history makes sense to increase the sampling to close to 100% for the next 120 days based on previous sampling program history."*

Reviewer 2: *"The analysis of seasonality is well done, at least given the limited dataset available, and the Chi square analysis clearly shows a statistically significant difference in one of the clusters of months that should be reflected in the final methodology."*
"…the authors have identified what I believe to be the most significant practices that might influence hazard and have identified a good pool of institutions to participate in any subjective encoding needed to determine the influence of these practices on hazard."

Reviewer 3: *"The proposed weights for the volume and risk score (last positive test, to be more precise) is consistent with the stated objectives."*

Reviewer 4: ***"****The method to assign a weight to sample history is straightforward and well justified."*

Reviewer 5: ***"****The weighting of seasonality and sampling history is based on individual odds ratio estimates and is consistent with the stated objectives."*

Charge 5 Critiques and Responses

Reviewer 1: NA

Reviewer 2: "*… the document does not explain why the Hazard Score is as it is: a 5 if there has been a positive sample and a 1 if not…. I suppose it was in some way related to the odds ratio analysis conducted, which showed that facilities with a prior contaminated sample were 2 to3 times more likely to have a subsequent contaminated sample*."

Response: Table 9 in the report shows that the odds ratio for plants that have had a positive sample in the last 120 days is app. 4.9 (95% confidence interval 3.2-7.5).

"*I can see the weighting employed for the production volume and hazard score using the 2 factors incorporated so far, but I don't see a description of how the weighting will be accomplished for the other two factors, and these factors are not to be found in the algorithm as yet. The text suggests that the algorithm will instead be adjusted eventually to include them."*

Response: Please see response to Reviewer 1 under charge 4.

Reviewer 3: "*It should be noted that while the report suggested the season and establishment's practice as factors for assigning the sampling weights, these are not currently incorporated into the algorithm and there are no specific formula details on how the weights will be determined. The intention of using these to further determine the final relative sampling weights is right but how the appropriate weights that combines these factors should be calculated is an important question and remain to be unclear. Some simulation studies may be needed in the future to help identifying the weights that will yield efficient and sensitive sampling designs."*

Response: Please see response to Reviewer 1 under charge 4.

"…the current algorithm assigns a rather arbitrary multiplicative factor of 5 if an establishment has been tested for E coli positive in the last 4 months, compared with a multiplicative factor of 1 for those tested negative."

Response: Table 9 in the report shows that the odds ratio for plants that have had a positive sample in the last 120 days is app. 4.9 (95% confidence interval 3.2-7.5).

Reviewer 4: **"***An explanation could be added why the model assigns an increased hazard score for all establishments positive within 150 days before sampling, whereas the report describes an increased risk within 120 days after a positive sample."*

Response: The algorithm is designed to assign samples fat the beginning of each month (app. 30 day window). Therefore, the first positive sample will be in data input to the algorithm from the preceding 30-day period.

"Strictly speaking, the relative risk or risk ratio (RR) would be a more appropriate measure for cross sectional data than the Odds ratio (OR). However, because E. coli O157:H7 contamination is a rare event, both measures are almost identical (OR=4.86; RR=4.79). The effect of season could be described accordingly. The RR for April-October compared to November-March would be 2.14 (95% CI 1.47-3.11), the OR 2.14 (95% CI 1.47-3.12)."

Response: As noted by the reviewer, the RR approaches the OR asymptotically for extremely rare events such as O157 positive samples. For the sake of consistency, we have reported all ratios as Odds Ratios in the revised report. Please see response to Reviewer 5 Charge 2 for more details.

"Different risk factors could be combined to a common hazard score by two different methods (1) multiplication of the RR or OR values for different risk factors (i.e. an establishment with a history of contamination in the season April-October would be assigned 10 hazard points, an establishment with no history of contamination 2 points); (2) calculation of OR in a multiple logistic regression model. The second method would be preferable, because the estimates for each risk factor could be corrected for the other risk factors. However, with this method risk factor information from scientific literature cannot be included. Estimating the effect of establishment practices will greatly rely on literature, because obtaining this information retrospectively may result in biased data."

Response: FSIS will consider the use of a logistical regression model when incorporating the establishment practices into the Hazard Score. Given the rarity of detectible O157:H7 contamination events and the complexity of the plant establishment data, a more effective method may be the latter one mentioned by Reviewer 4 — namely the use of scientific literature that provides quantitative data on the level of O157 reduction. As mentioned previously in this document, FSIS will have the methods and revised algorithm peer reviewed once the establishment practices are incorporated. (For more details, see the report and response to Reviewer 1 Charge 4).

"*The way in which production volume is currently handled in the algorithm has several disadvantages. First, production volume has an effect on the exposure of consumers as well as on the risk of an establishment to produce contaminated meat. According to the data provided in the report, establishments in category 3 have a 2 times greater risk of positive samples compared to category 4 establishments (RR=2.34, OR=2.35). This information is currently not used in the algorithm. It could be argued that this risk factor information should be included in the hazard score rather than in the volume score, because the latter reflects exposure of consumers. Second, volume score is recorded on a flexible scale (1-4, 2-5 or 3-6) to fulfill the decision criteria stated in the report. If the higher scale is used, the relative weight of volume score compared to hazard score is greater than if the lower scale is used (production volume is more important in driving sampling probability if the higher scale is used). Some kind of standardization should be applied to avoid this problem (either choose a large enough scale to ensure that the criteria can be met with a fixed value for the highest category, or allow decimal numbers in the score) …* I*t would be beneficial to achieve more*

transparency on how the relative importance of volume score and hazard score is determined. One option would be to standardize the production volume point values and the hazard point values to a scale with a common maximum and minimum value before multiplying them to obtain the plant probability points. Another (somewhat more complicated) option would be to define a fixed relative importance of volume score and hazard score in determining sampling probability. In this case, the volume score points would be corrected by multiplying its value for each establishment with a correction factor X, which depends on the relation of the sum of all volume points V_i to the sum of all hazard points H_i, where Y and Z describe the relative importance of volume score and hazard score in determining sampling probability (e.g. Y=40%, Z=60%)"

$$X = \frac{Y \cdot \sum_{i=1}^{n} H_i}{Z \cdot \sum_{i=1}^{n} V_i}$$

Response: In response to this comment, and others throughout the peer review, we have revised and greatly simplified the volume score calculations in the algorithm. Please see the report and response to Reviewer 2 Charge 2 for more details.

CHARGE 6

Comment on the performance measures described in the report. Do they provide an objective measure of the program's performance? Why or why not? If not, provide alternative performance measures. Please keep in mind that the performance measures must be quantitative, objective and based on data readily obtainable by FSIS.

- *i) How effectively testing resources are utilized to monitor the greatest percentage of beef products generated.*
- *ii) The ratio of the prevalence of positives in the risk-based sample pool to the prevalence of positives in an unweighted random pool.*"
- *iii) If the program is verifying the safety of a (i) greater portion of product and (ii) the riskiest portion of product, then it is reducing the exposure of consumers to E. coli O157-contaminated ground beef (see above measures).*
- *iv) The number of human illnesses directly prevented by detection of an E. coli O157:H7-positive lot by the program.*
- *v) Use surveillance data from the Centers for Disease Control and Prevention (CDC) to estimate the number of E. coli O157:H7 illnesses due to ground beef*

Charge 6 Commendations

Reviewer 1: "*Estimating the percentage of the supply verified each month by dividing total pound produced by pounds verified is one way that the Agency can report on the performance of the program."*

Reviewer 2: i) "*This is both a completely reasonable objective, being public health protective, and one that the proposed methodology achieves."*
ii) "This is a very good metric, and the proposal as to how it would be determined is sound"
iii) "This is another good measure and, as the authors state, this measure is assured by the first two measures above."

Reviewer 3: "*The performance measures for Objective 2 and 3 are objective and convincing. There is clearly an increased proportion of the establishments with high production volumes being tested under the proposed scheme (Table 6 – Table 10)."*

Reviewer 4: **"***Both measures suggested to evaluate how effectively resources are being used are valid and provide good information on different aspects of the performance of the program."*

Reviewer 5: **"***Two categories of performance measures are proposed. Both are reasonable albeit with some limitations."*

Charge 6 Critiques and Responses

Reviewer 1: "*Other measures such as comparing the weighted vs unweighted results may not be useful because of the low prevalence of E. coli O157:H7 in the FSIS samples. Measuring the public health impact of the program will probably difficult because of the reasons stated that there is too much uncertainty in predicting illness based on the incidence of positive samples or surveillance data."*

Response: We agree with the reviewer. FSIS is currently working on improved methods for measuring the effectiveness of its risk-based programs. *E. coli O157:H7* programs face a particularly difficult challenge due to the rare occurrence—but potentially severe health outcomes-- of the pathogen.

Reviewer 2: ii) *"I share the authors concern, however, that the low occurrence rates of contamination will make the comparison statistically unreliable, at least until several years of data are collected"*

Response: We agree with the reviewer. FSIS is currently working on improved methods for measuring the effectiveness of its risk-based programs. *E. coli O157:H7* programs face a particularly difficult challenge due to the rare occurrence—but potentially severe health outcomes-- of the pathogen.

iv) *" It also relies quite heavily on being able to track disease outbreaks to a specific exposure pathway such as beef, unless the authors mean to use an exposure-response relationship to estimate disease incidence rather than measuring incidence directly (as in the next bullet). The difficulty in the former approach is that it is necessary to not only predict exposure-response functions (which are not yet well developed in the literature), but to also model the intervening influence of food transport, storage, preparation and consumption. There are significant uncertainties associated with these steps, particularly due to the re-growth of the microbes as beef is stored."*
v) "I am the most skeptical of this approach, although it is the one that would provide the most direct evidence. It is difficult to ascribe outbreaks to a single exposure pathway, and there is the large problem of under-reporting of disease"

Response: We agree with the reviewer about the difficulties inherent in human illness based performance measures. FSIS is currently working with the CDC to improve our models for attributing human illness to particular products. It will be important to build models that account for the uncertainty of outbreak detection, disease under reporting and accurate attribution to complex food vehicles.

Reviewer 3: *"… quantitative measure of performance regarding Objective 1 is not demonstrated. The report suggested a good measure of effectiveness (4th paragraph, p.30) but does not provide any results, demonstrating that the new scheme indeed catches more positive samples than the existing on. This can be addressed with a simulation study where one mimics the parameters in the real situation and uses both algorithms to sample a fixed number of test samples*."

Response: We agree that a simulation approach may be an appropriate solution given the other difficulties inherent in measuring the changes to a program with such low levels of positive samples. In response, we have built a bootstrap simulation model that re-samples the risk-based results to simulate a simple random sampling of the frame. The model is still currently under a testing and refinement stage.

Reviewer 4: *"Running a random and a risk-based sampling program in parallel would be a huge effort. An alternative could be a stochastic simulation which compares the effectiveness of both sampling strategies, while taking the uncertainty of risk factor and prevalence information into account. An additional useful measure could be the cost associated with detection of one contaminated sample. Measuring the public health impact of the sampling program will be quite difficult. The greatest public health impact of the improved surveillance program will probably be through prevention of contamination of ground beef and beef trim by sensitizing producers to the risk of E. coli O157:H7, and by providing incentives for good manufacturing practices. Thus, the*

measures suggested here are likely to underestimate the effect of the sampling program on public health."

Response: We agree with the reviewer. For comments on the simulation approach, please see the above response to Reviewer 3. The problem of measuring the indirect effects (such as improved establishment practices) of the sampling program on O157:H7 levels in the ground beef supply is a difficult one. We also agree that these are likely top be the largest impacts of the program and thus the types of direct performance measures discussed here will be a large underestimate of the overall impact. One possible approach to addressing this concern is a longer-term modeling of the trends in O157 illness caused by ground beef (see response above to Reviewer 2).

Reviewer 5: ***"****…the percentage of supply verified each month will be estimated and second, the portion of potentially contaminated product that is verified will also be estimated. As the authors indicate, a random, unweighted sampling program needs to be run side-by-side with the algorithm and results of the 2 programs directly compared. Decisions relative to whether this is a viable choice would need to be based on cost and likely sample sizes to detect meaningful differences in prevalence."*

"The second measure, the public health impact of the sampling program, is much more difficult to quantify because of the number of intervening factors between reduced product contamination and a lower burden of human illness. The goal is laudable and one with which we collectively struggle. The authors justify the program as reducing the exposure of consumers to E. coli O157-contaminated ground beef. Theoretically this should be true, but without a side-by-side evaluation of the risk-based and the traditional program, this contention it is purely speculative. The second choice (prediction of the number of human illnesses directly prevented by detection of an E. coli O157:H7-positive lot by the program) is vague. Does the current risk model for O157:H7 allow this to be done readily? The use of CDC surveillance data is reasonable but a limitation now especially for sporadic E. coli illness is that consumers have become more aware of alternate sources of E. coli contamination and might be less likely to report an exposure to ground beef."

Response: We agree with the difficulties described by Reviewer 5 and appreciate their support of our efforts with this difficult undertaking. Please see response above to reviewer 2 for comments on the use of outbreak data and human illness modeling.

ADDITIONAL REMARKS

Reviewer 1: NA

Reviewer 2: NA

Reviewer 3: NA

Reviewer 4: Table 2 of the report has incorrect headings. Instead of "% Meat" the heading should read proportion (or the numbers changed to a % format). Also, you might want to explain that the proportion of trim is referenced towards the meat, not the total carcass weight.

Response: Suggested changes made to Table 2.

The excel file 'Testing data' contains 4 invalid dates in the column 'collect date' (1900 instead of 2000 and 1901 instead of 2001). This error should not have an effect on the analyses, though, because the column 'analysis end date' with corrected dates is used instead of 'collect date'.

Reviewer 5: P4 – the term "poor performers" is used here and again on page 34 – for clarity, a definition (or an example) of a 'poor performing establishment" should really be given in the document.

Response: The report has now been modified to define "poor performers" as the following "*defined in the initial phase of the algorithm as establishments that have tested positive for O157:H7 in the past 4 months and expanded to include plants with high risk practices in the future version of the algorithm*."

P5 – hyphenate "high-risk" when used as an adjective – same comment applies on page 22.

Response: Changes made

P5 – last sentence – suggest change to "… should therefore increase the public health impact through more efficient allocation of FSIS resources."

Response: Change made

P6 – superscript for reference 5 to 9 is missing - were these references used?

Response: References changed

P7 – Isn't it true that other sources, e.g. produce, are currently as least as important and under increased surveillance for contamination? It would helpful to incorporate food attribution data, but I assume these are not available.

Response: We estimate that app. 40% of O157:H7 illnesses are attributed to beef products. Attribution data and a discussion of attribution modeling are outside the scope of this particular report.

P7 – references 20 to 23 are not in the reference list.

Response: References removed.

P8 – why was the last 4 months chosen as the time window for sample history?

Response: This is explained in detail elsewhere in the report ("sample history pp14-22)

P9- define "high prevalence season" here, or at least refer the reader to section where the seasonal data are evaluated (pages 19 to 22).

Response: We have added a referral to the report section that describes the seasonal analysis and definition of the high prevalence season.

P9 - The sentence "The risk-based sampling program will be more representative of the beef supply than the current sampling program that provides random sampling by establishment (i.e., without consideration of the amount of product produced, interventions, sampling history, etc.)" is not really correct. Only the first point "without consideration of the amount of product produced" is relevant to beef supply per se.

Response: We have changed the wording accordingly.

P9 and throughout the document – what was the justification for not using the most recent data, i.e. 2006, in the analyses?

Response: At the time, the algorithm was developed and the report was written the analyses used the most current data available.

P11 – why is there a need to assume an "upper bound of 500,000 lbs" for category 1? Aren't there data relative to production levels for these establishments?

Response: FSIS collects production volume data by an open-ended scale (i.e. where the upper end is > 250k lbs/day. We have estimated the average production volume of these largest plants to be 375k/day by assuming an upper limit of 500k for this group. The 500k limit is based on the opinion of people familiar with the industry such as FSIS inspectors.

P11, Table 2 – if the column headings (% meat and % trim) are correct, then the respective values should be 70,70,70,70 and 18,18,53,90.

Response: Changes made to Table 2.

P11 – define "MT03" samples here since the acronym is used here for the first time in the document. This could be readily achieved by copying or moving the relevant definition which appears in the last paragraph on page 14.

Response: Revision made.

P12, Table 4 – different numbers of establishments are used in this table compared with other places in the document -- here 1668 versus 1536 (previously) and 1540 (later). Perhaps a single sentence could be added somewhere to account for the differences.

Response: The differences are related to the years being analyzed since new establishments arise and existing ones leave the industry, giving rise to fluctuating numbers.

P12, Table 5 – change "no cat" to "not available" or "N/A" to be consistent with usage on page 11.

Response: Revision made.

P12, Tables 4 and 5 - remove % in 2 right-most columns because this is in the table column headers already (i.e., % total production volume, % total samples).

Response: Revision made.

P17 – wording of first sentence below the table should be change to reflect the comparison group. New wording suggestion – "Thus, random samples collected within 120 days of a previous positive sample at the same establishment are more likely to be positive for *E. coli* O157:H7 than samples collected within 120 days of a previous negative sample" . This assumes that the comparison categories will be mutually exclusive.

Response: Revision made.

P14 to 18 – It would be helpful to have some background information on sampling methods, numbers of laboratories doing the testing and variability, if any, in *E. coli* test protocols. A question that might arise is whether some establishments have more positive results just because their sampling and testing protocols are more rigorous and sensitive than protocols used at other establishments.

Response: The results analyzed here are all from FSIS laboratories that use a standardized protocol for sample analysis. Method validation is used to minimize variability from lab to lab. In addition, samples from any given establishment are not dedicated to a single FSIS lab but are distributed among the three facilities.

P23, L2 – should be "prevalence", not "incidence" since it is snapshot picture in time.

Response: Revision made.

P24, Table 16 – "does test" not "does tests".

Response: Revision made.

P25, L4-5 – this sentence implies that the survey has been done. Is this correct?

Response: The survey was administered for the first time in December 2007. The report has been revised to state this.

P26, second point – the reality of slaughter plants is that there are often a series of sequential mitigations to reduce the frequency of *E. coli*. How will the joint effect of mitigations be considered in this evaluation?

Response: Although they are important for the future of the algorithm, those considerations are outside the scope of the current report. As described elsewhere in this document, the report and algorithm will be peer reviewed for a second time once the establishment practices are incorporated.

P27, last 2 lines – A brief explanation (or a cross-referencing to Table 9 on page 16) should be given to justify the choice of hazard scores.

Response: The explanation and reference have been added to the report.

P28, Figure 2 – why do the production volume box and the O157 result box both have the same designator, M2K?

Response: This is a reference to an internal FSIS database that contains the relevant data

P29, last line – should be "data are".

Response: Revision made.

P30 – I question the correctness of the statement "The overarching goal of FSIS' *E. coli* O157:H7 testing program is to help ensure that industry is producing raw ground beef that is free from O157:H7 contamination". Shouldn't it be minimal risk or a similar modification of the wording? Given the small sample that is collected and tested, and the imperfect sensitivity of testing methods, the statement as written is misleading.

Response: We have revised the statement accordingly. For the sake of clarity, however, FSIS has never meant to imply that its testing program ensures that all ground beef is free from O157. Only that the goal of the program is to ensure that the establishments that produce ground beef are controlling the O157 pathogen to the best degree possible.

P31, Table 19 heading – hyphenate "risk-based".

Response: Revision made.

P31, Table 19 footnote – "unweighted" not "unwweighted". Also last sentence seems incomplete. Perhaps it should read "In the proposed risk-based weighted sampling program, an estimated 8,000 annual samples….."

Response: Revision made.

P34 – suggest deletion of sentence "Establishments with recent *E. coli* O157:H7-positives will be sampled monthly for four months" since this is at odds with the last sentence in the same paragraph and with statements made earlier in the document about no establishment being sampled with probability of 0 or 1.

Response: Revision made.

P34 – why 1540 ground samples when 1536 were presented in Table 1?

Response: We had rounded the number 1536 to 1540 in the text on page 34. We have revised the text to read the more precise number of 1536.

P35 – how will a "good testing program" be defined?

Response: "Effective" testing program is a more precise definition. We are currently assessing the survey data recently collected that includes data on the frequency, volume, and methodology of plant sampling programs. Once we have analyzed this data we will be better able to rank the establishments testing programs. As stated elsewhere, once the algorithm and report are revised with these additions there will be a second peer review.

Appendix A: Reviewer Biographies

Douglas Crawford-Brown. Dr. Crawford-Brown is Professor in Environmental Sciences and Engineering and in Public Policy, and Director of the campus-wide Institute for the Environment, at the University of North Carolina at Chapel Hill. He received his degrees in physics (BS, 1975; MS, 1977) and nuclear science (PhD, 1980) from the Georgia Institute of Technology. His activities focus on the modeling of human health risks – primarily of carcinogens and microbes -, modeling of alternative policies to tackle a range of environmental problems, and development of tools of risk assessment for application in risk-cost-benefit assessments and uncertainty analyses. He is the author of 130 academic articles and 5 books on these topics and has served on a wide variety of state, national and international commissions addressing environmental issues. These include Federal Advisory Committees for the EPA on Endocrine Disruptors, the National Pollution Prevention and Toxics Advisory Committee, the National Drinking Water Advisory Committee (CCL subgroup) and the Clean Air Scientific Advisory Committee. He has developed dose-response models for a wide variety of microbial contaminants in food and water, and applied these models in assessments of the risks of food-borne contamination and the identification of strategic points of intervention for food safety.

Philip H. Elliott. Dr. Elliott is currently Director, Microbiology for the Grocery Manufacturers/Food Products Association in Washington, DC. He manages a staff of microbiologists and thermal process authorities in providing technical services and research projects for the Association's members. He has been an instructor several HACCP training and allergen control courses for the Association. Previously he was the Director of Quality Assurance for the Pinnacle Foods Corporation, Manager of Food Safety and Microbiology for Vlasic Foods International. He managed the microbiology group for the Campbell Soup Company and was Senior Microbiologist for Armour-Dial. He has a Ph.D. from Rutgers University in Food Science. His area of research was developing predictive models for growth and toxigensis of nonproteolytic *C. botulinum*. He has a B.A. in Biology and M.S. in Food Science from the University of Delaware.

Ian Gardner. Dr. Gardner is a Professor of Epidemiology in the Department of Medicine and Epidemiology at the University of California, Davis. He received his B.V.Sc. from the University of Sydney, Sydney, Australia; his M.P.V.M. in Epidemiology and Ph.D. in Comparative Pathology from the University of California, Davis. His research expertise includes: risk analysis related to livestock health and food safety; diagnostic test evaluation; epidemiology of infectious diseases in livestock production systems; epidemiology of protozoal myeloencephalitis in marine mammals and equids and epidemiology of catastrophic musculoskeletal injuries in racehorses.

Gertraud Regula. Dr. Regula is a veterinary epidemiologist at the Federal Veterinary Office of Switzerland. She is leading a food safety research unit which works on applied research on zoonoses and antimicrobial resistance. Her research interest is improving the efficacy of monitoring and surveillance programs. She has been involved in the

development of simulation models for the evaluation of risk-based sampling strategies in the monitoring of residues and antimicrobial resistance. Dr. Regula is a lecturer at the University of Bern, and a Diplomate of the European College of Veterinary Public Health.

Haibo Zhou. Dr. Zhou is Associate Professor at the Dept. of Biostatistics, University of North Carolina at Chapel Hill. He is the Director for Biostatistics for the Center for Environmental Medicine, Asthma, and Lung Biology at UNC. He collaborates with investigators at National Institute of Environmental Health (NIEHS) and the U.S. EPA Human Study Division. His statistical expertise is in outcome-dependent sampling, survival analysis, missing data and auxiliary data problems. Dr. Zhou is interested in environmental statistics, reproductive epidemiology, human fertility, children's health development, risk assessment, and respiratory diseases due to environmental exposures such as smoking and air. He has published extensively in both statistical journals and the subject matter journals. He was the PI on two NIH R01 grants and served on NIH grant review panels. He is currently an associate editor for Biometrics, a leading professional journal in statistics. Dr. Zhou holds a Ph.D. and M.S. in statistics from the University of Washington.

Appendix B: Peer Review Charges

The "charge to peer reviewers", as defined in the OMB's Peer Review Guidelines, consists of the issues and areas we would like you to focus on in your evaluation of the risk assessment (report, analysis, and model). The charge to the peer reviewers for this risk assessment evaluation follows. Please address each question or issue:

a. Is the overall approach for allocating samples through a probabilistic draw sound? (a random draw where the probability of selection is weighted by risk) If not, what problems exist and how should they be addressed?

b. Evaluate algorithm source code and mathematics.
    1) Are the techniques (mathematics and equations) appropriate? If not, provide alternatives.
    2) The reviewer should examine and verify that the data analysis and source code are accurate.

c. Does the algorithm accomplish the three objectives (described in the report)? Please be specific. If so describe why and if not suggest how it could be altered to better achieve the described objectives.

d. The algorithm uses four general areas to assign the overall risk of causing O157:H7 illness (production volume, sample history, season and establishment practices). Comment specifically on the use of each of these areas. Are there additional factors that should be considered?

e. Each of the four areas (production volume, sample history, season and establishment practices) is assigned a weight in the algorithm to determine the overall probability of sampling for each establishment. In some cases weighting is proportionate (i.e. seasonality and sample history are weighted proportionate to the increased probability of detecting contaminated lots) while in other cases the weighting needs to be balanced with the overall objectives of the program (i.e. production volume weighting is more complex due to additional sampling concerns). Comment specifically on the method of weighting each area. Is the weighting consistent with the three objectives?

f. Comment on the performance measures described in the report. Do they provide an objective measure of the program's performance? Why or why not? If not, provide alternative performance measures. Please keep in mind that the performance measures must be quantitative, objective and based on data readily obtainable by FSIS.

Appendix C: Complete Reviews

Reviewer 1:

SUBJECT: Review of FSIS "Risk-Based Sampling for *Escherichia coli* 0157:H7 in Ground Beef and Beef Trim"

I find the approach of using a random draw with weighted probabilities for selection of plant to be sampled is an appropriate method to meet the objectives of this program and the Agency. This addresses the issue of using Agency resources to best possible advantage to find sources of *E. coli* O157:H7 in ground beef and trim. The proposed method of weighting the probability for selection currently relies on volume and sample history, but as more information is gathered on establishment practices that will effectively reduce the incidence of *E. coli* O157:H7 on beef, I feel that this algorithm could be a very powerful tool in focusing the testing efforts of the Agency.

Although samples of beef are currently found to be positive for *E. coli* O157:H7, taking 8,000 samples per year to monitor 3.7 billion pounds of trim is really like looking for a needle in the proverbial haystack. If uniformly applied based on volume alone that is something like each sample representing approximately 500,000 pounds of trim. It is commonly recognized that product testing is not the way to control the hazard but at least this verification testing keeps the industry honest in their attempts to prevent the organism from being in the product.

The mathematical approach for this program seems to appropriate for making risk based decisions. I am not qualified to comment on the source code but in trying out the model the program and the results made sense to me.

The algorithm does aid in increasing the proportion of samples taken at establishments that are more likely to produce product contaminated with *E. coli* O157:H7. This will improve as the other factors like seasonality and establishment practices are included in the model. It meets the objective of allocating Agency resources more efficiently and verifies all eligible establishments at a reasonable frequency.

Although the objective of using the risk based algorithm is to more fairly distribute the sampling where it would be useful, if looking at volume alone, the smaller plants (<1,000 lbs./day) are still being sampled more frequently on a per pound basis than are the larger plants. For example using the information in Table 19 and in Figures 6-10, if a small plant makes the maximum of 1,000 lbs / day, they would produce approximately 20,000 lbs /month. If a large plant made the minimum of 250,000 lbs./day, they would be making 5,000,000 lbs / month. With the current system, if a small plant was sampled an average of 7 times per year, each sample would represent ~34,000 lbs of product. If the large plant was sampled 7 times, each sample would represent ~8.5 million pounds of product. Using the algorithm to assign probability based on volume alone, the small plant would be sampled an average of 4 times with each sample representing 60,000 lbs of product. The large plant would be sampled 11 times with each sample representing 5.5 million pounds of product. This is going in the right direction of being equitable

sampling, but is still a long way off. I understand that these sampling rates cannot be equal and still allow for a meaningful sampling of the small plant. In the scenario above, it would take the small plant more than 20 years to make as much product as the large plant makes in one month. That is a long time between samplings! As the other risk factors of history, seasonality and establishment practices are included, the algorithm will give a better recommendation in sampling based on the actual probability of finding *E. coli* O157:H7. Another risk factor to include might be geographic location of the plant or source of the cattle.

The weighting of the volume is balanced with the overall objectives of the program. This is consistent with the objectives of the program to verify all eligible establishments at a reasonable frequency. The weighting of sample history makes sense to increase the sampling to close to 100% for the next 120 days based on previous sampling program history. Seasonality seems to be an important factor that may be included sooner rather than later in the algorithm as soon as the Agency can get a handle on how that will affect their laboratory resources. It will be interesting to see how the establishment practices are weighted in the future since they may have the greatest potential for predicting the probability of finding a positive sample.

Estimating the percentage of the supply verified each month by dividing total pound produced by pounds verified is one way that the Agency can report on the performance of the program. Although, for the reasons described above, more of the smaller plants' product still will be sampled, assuming all other risk factors being equal. Other measures such as comparing the weighted vs unweighted results may not be useful because of the low prevalence of *E. coli* O157:H7 in the FSIS samples. Measuring the public health impact of the program will probably difficult because of the reasons stated that there is too much uncertainty in predicting illness based on the incidence of positive samples or surveillance data.

The Agency should be commended for taking this approach in sampling ground beef and trim in effort to reduce the incidence of *E. coli* O157:H7 in the nation's food supply as well as making better use of their resources to monitor food safety.

Reviewer 2:
Review of the FSIS Methodology for Sampling

Question 1: *Is the overall approach for allocating samples through a probabilistic draw sound? If not, what problems exist and how should they be addressed?*

The general approach is sound overall. This kind of sampling is quite common in both scientific and regulatory practice, and so most users of the method should find it understandable. Embedding the approach inside a Monte Carlo routine is the correct thing to do, and the authors have correctly executed (in the code) the Monte Carlo procedure.

The overall approach is a form of Analytic Hierarchy Process, which also has a long history of use in decision-making. In this regard, the authors have laid out clearly the specific attributes they will consider (they mention 4 general categories of attributes, although they have implemented only 2 to date); have specified a procedure for quantifying each attribute; have specified a scale for reducing these quantitative values to a scale (of 1 to 4 or 1 to 5); and have specified an algorithm for combining them into a single score used to specify the probability intervals for each facility. In all of these steps, they have used well accepted methods.

The only quibble I would have at this very general level is their choice of a multiplicative function for the 2 attributes they consider (volume and past results). There is nothing wrong with that algorithm per se, and it does have the good feature of giving each of the two attributes roughly equal weight in the sampling. But there are other algorithms they might have used, such as specifying a measure of "importance" to each of the two attributes and then taking a weighted sum (score from attribute 1 times the importance of attribute 1 + score from attribute 2 times the importance of attribute 2 + …). I am not at all suggesting they change the algorithm, but it would be helpful to the reader to understand why they chose this particular multiplicative function. The form of the function does, in the end, affect the sampling probabilities assigned to each facility.

Continuing with this issue, the document does not explain why the Hazard Score is as it is: a 5 if there has been a positive sample and a 1 if not. I realize the desire is to have a score that is roughly equivalent in magnitude to the Volume Score, but there is a need to explain more clearly why the particular scoring used was developed. I don't think it is a bad scoring system for Hazard, but I kept looking for a place in the document where the choice was justified. I suppose it was in some way related to the odds ratio analysis conducted, which showed that facilities with a prior contaminated sample were 2 to3 times more likely to have a subsequent contaminated sample.

On a final point, their choice to sample without replacement is the correct one.

Let me suggest that the authors include a much improved flow diagram of the computational steps, showing the decision structure. This would walk the reader step-by-step through the process followed in the code. The code could then use Comment lines to identify where a specific section of code executed a specific step in the computational flow diagram. I found it extraordinarily difficult to move between the text and the code, and many aspects of the code, especially in the area of the Volume Score, are inadequately described in the document. A full flow diagram, showing the steps of the algorithm and decision points, would alleviate this problem.

Question 2. *Evaluate the algorithm source code and mathematics. Are the techniques (mathematics and equations) appropriate? If not, provide alternatives. Examine and verify that the data analysis and source code are accurate.*

This part of the review was by far the most difficult because the document does not properly lay out the mathematical details or the decision process. I had to go through the

code line-by-line to find the answers to many questions that should have been answered in the document. This was particularly true for the Volume Score (it was easy to find the answers for the Hazard Score, both in the document and in the code, since it is relatively straightforward).

The main problem with the Volume Score lies in the reader determining whether the four criteria shown in the box diagrams are intended to apply to individual facilities and their probability ranges (the ranges used in the Monte Carlo selection process) or only to volume categories. This is never explained in the document, and so I went into the code expecting to find the criteria applied in some way to the individual probability ranges. After going through the code, I came away convinced that the 4 decision criteria are being used to determine (or bound) the NUMBER of samples to be selected from within each volume category, and NOT the probability range applied to any specific facility. If I am not correct in this, then there is something about the code that is eluding me. And that is problematic, because I have written, and reviewed, many codes of this type, and if I am confused I can only imagine what the average user of the code will make of the computational structure and details.

This surely is a problem of the documentation, which is need of great improvement. Both of the figures that show the decision framework (Figures 2 and 5) suffer from the same deficiency. Based solely on the document, it is impossible for the reader to understand how the Volume Score is calculated (the same is not true of the Hazard Score). This then makes it impossible for the reader to confirm the results in Table 18. There must be a complete set of calculations for at least a few example facilities (establishments or sites) showing exactly how the probability ranges were determined. It is clear from the Table that the probability range associated with a given facility is related one-to-one to the Probability Points in the first column. All facilities with a Probability Points of 4 have an interval width of 0.00108; all with a Probability Points of 2 have half of this width; etc. That much is clear and fully defensible, and the code executes this properly. But as a reader, I had to find this pattern myself. The document should have explained it clearly.

A minor quibble is with the way the Volume Score needs to reverse the order of the volume categories in starting with the initial score. This is uncovered at a point in the code when a calculation of 5 – i is performed, with i being the category number. If a 4 had originally been assigned to the highest volume category (> 250,000 pounds) and a 1 to the lowest volume category (rather than the reverse), this step of calculating 5 – i would not be needed. I can see no reason why the approach used in the code was taken, as it adds nothing and just creates another potential point of confusion for the reader.

Table 18 also will cause confusion for another reason. The document speaks of the algorithm in which there is a Sampling Probability, p_i, which is used as the basis for drawing samples. The reader's attention is not drawn to the issue of Probability Points, and so the relationship between these Points and the Sampling Probability is not clear. I spent many hours trying to track this down through the code, hours that would not have been needed if the document were better written with step-by-step computational details. My conclusion here is that the document does a good job of describing the qualitative

aspects of the procedure, and the code executes what I finally determined to be the computational steps, but the document is completely inadequate in linking the qualitative discussion to the computational steps, and the computational steps to the code. The writers should bear in mind that trying to learn the computational steps by unpacking someone else's code is almost impossible, especially since the code does no provide enough Comment lines to make the reasoning clear.

Given this lack of clarity, I was not able to reproduce Figures 6 through 10. They are very interesting figures, and certainly support the claim that the procedure developed here does a better job of sampling based on the potential for exposure and risk than does the existing method of sampling all facilities equally. But absent better documentation, I cannot reproduce the figures and so I cannot vouch for the validity of their results. I am not saying they are incorrect, only that I could not verify them despite trying.

Question 3. *Does the algorithm accomplish the three objectives? Please be specific. If so describe why and if not suggest how it could be altered to better achieve the described objectives.*

I take these three primary objectives to be:

- *To increase the proportion of FSIS samples taken at establishments that are more likely to produce product contaminated with E. coli O157:H7*. My answer here is that the methodology does indeed accomplish this. This is evident both in the structure of the methodology and in the results in Figures 6 through 10. The risk "metric" here is something akin to probability of contamination times volume of product. If we take this as a reasonable metric of public health risk (and I think that this is a valid assumption), then this procedure will do a better job of sampling facilities with a probability that is related to this exposure or risk. My only quibble here is that the Hazard Score is not actually proportional to the extent of contamination. It is instead a measure of the occurrence of contamination rather than the concentration. But even this quibble disappears if the assumption is made that any measurable contamination in the product emerging from a facility will at some point result in the growth of the microbes, leading to an exposure that exceeds a threshold for disease. I presume this is what the authors intend.

- *To allocate FSIS resources more efficiently by verifying a greater portion of the U.S. trim and ground beef supply with the same number of samples as the current program*. My answer here is in the same as in the first objective. The methodology clearly produces a more efficient sampling scheme than that used currently, in the sense of allocating limited sampling resources to target the points of most effective intervention in the food supply. I am not, however, fully convinced that the first criterion in bounding the Volume Score, the one related to fraction of establishments and fraction of the total

supply, is needed. I understand the desire to have at least some establishments in each category sampled, and the danger that NONE of the establishments in the lowest production category might be sampled if a strictly probabilistic sample were drawn. But these two boundaries don't seem to me particularly justified given the desire to move towards the more probabilistic approach. This objection might be overcome by adding a paragraph describing the implications of this criterion on the sampling, perhaps showing an example of the sites selected when the criterion is not employed and when it is added in. Presumably, with the criterion in place, there will be a shift in sampled sites towards sites in the categories that have a smaller percentage of the total production volume. An example of this would be useful in the document.

- *To verify all eligible establishments at a reasonable frequency regardless of an establishment's production volume, interventions, or predicted public health risk associated with their product*. This part of the approach meets the goal entirely through a bounding or constraint that requires each site to be monitored at least once in some prescribed period. This was, however, one place where I could not fully understand how the code was accomplishing this part of the approach. That needs to be explained more fully in the document.

Question 4. *The algorithm uses four general areas to assign the overall risk of causing O157:H7 illness (production volume, sample history, season and establishment practices). Comment specifically on the use of each of these areas. Are there additional factors that should be considered?*

My comments here are framed by what I see as the central goal of the algorithm: to identify sampling frequencies based on some measure of risk to public health, meaning in this case a mathematical product of an exposure measure and a measure of the hazard posed by a volume of beef product. In this algorithm, the exposure measure is quantified by the volume of beef product and the hazard measure is quantified by some likelihood that a given volume of beef product is contaminated.

As I see it, the exposure measure is quantified entirely by the production volume. Given that one cannot say much about where the product will end up, I think this is probably the most precise measure one can use. And so I fully support the idea of using it to quantify v in the probabilistic sampling calculations.

The authors then intend for three factors to go into the measure of hazard: (1) past sampling results, which provide an indication as to whether a particular facility has had a problem with contamination in the past; (2) seasonality, since there is good evidence that occurrence of contamination is dependent on season; and (3) introduction of practices that might reasonably be expected to reduce the occurrence of contamination (or whose absence will increase the occurrence).

These seem to me three very reasonable factors to include. The second factor requires stratifying all of the contamination results by season, but the authors have already done this using the current data base and the results suggesting a seasonal pattern are compelling. As a result, they are on firm ground in using this seasonal pattern to adjust the final Sampling Probability.

It also is very reasonable to assume that practices can influence the occurrence rates of contamination. Here the problem will be that the database that might create the correlations between specific practices and the occurrence rate for contamination is not available (as far as I can tell). So it is not yet clear how they will develop these correlations. They might need to consider using a semi-quantitative measure of the effect of practices by using subjective judgment to estimate the likely impact of each practice on the likelihood of contamination and then assigning a scale from 1 to 4 (1 being no practice is in place, 4 being some sort of redundancy of practices). Ideally, though, the database need to develop the actual correlations will be available (I have no suggestions as to where one might go to obtain such a database, if it even is available). The authors have, however, identified a good group of institutions to participate in any program to assign these semi-quantitative measures, and have properly identified the practices this group should consider. Given the number of practices, however, I am not sure that the database will be sufficient to develop judgments for each practice individually, so clusters of practices might need to be considered.

What I am not at all clear on is how these additional two factors (items 2 and 3 above) will go into the complete algorithm. I believe they all fall under the Hazard Score, at least in principle, which suggests the three factors will be collapsed into one Hazard Score, to then be combined with the Volume Score in precisely the manner already executed. But the authors then face the difficult decision as to the manner, computationally, for folding these three factors together. A possibility is to develop a scale of 1 to 5 for each of the three factors (I have chosen 5 here because that is the scale used currently for occurrence), then multiply the three factors, and then rescale to a final 1 to 5 score for Hazard (presumably, a facility with 5 x 5 x 5 = 125 would then receive a 5, so a total product of 125 becomes the high end of the Hazard scale, which would then be equated with a final Hazard Score of 5, and a 1 x 1 x 1 would constitute the low end of the scale, or a final Hazard Score of 1.

An alternative to this is to use the occurrence data as the primary score (as is currently the case) and then make this score conditional on the answers to items 2 and 3. One such scheme would be to assign a facility with past contamination a score of 4 if there has been past contamination and a 2 if there has not, and then to add or subtract a point depending on the answers to the other two factors (e.g. a 4 would become a three if there are protective practices in place and a 5 if there are not).

Question 5, below, seems to imply that the algorithm has already been worked out for combining all 4 factors, since it is mentioned that each factor is “assigned a weight”. This decision isn’t evident to me in the document, which instead says that these additional 2

factors will be incorporated in 2008. Perhaps I missed something, but I cannot see in the document or in the coding where this algorithm has been specified.

Question 5. *Each of the four areas (production volume, sample history, season and establishment practices) is assigned a weight in the algorithm to determine the overall probability of sampling for each establishment. In some cases weighting is proportionate (i.e. seasonality and sample history are weighted proportionate to the increased probability of detecting contaminated lots) while in other cases the weighting needs to be balanced with the overall objectives of the program (i.e. production volume weighting is more complex due to additional sampling concerns). Comment specifically on the method of weighting each area. Is the weighting consistent with the three objectives?*

Let me start by saying that the Executive Summary says that "The risk-based *E. coli* O157:H7 sampling algorithm (described in this report) allocates samples in a random draw where the probability of each establishment being sampled is weighted by three factors: FSIS microbiological test results for *E. coli* O157:H7; an establishment's *E. coli* O157:H7 interventions and testing programs; and an establishment's production volume." So the reader is not alerted that there is the fourth factor, seasonality. This seems to suggest that seasonality is not so much a factor to be included with a separate weight, but perhaps a modifying influence on one of the Hazard Score factors that ARE provided a weight. And in any event, the current document and code seems only to reflect production volume and past sampling results.

As I mentioned in the previous question, I can see the weighting employed for the production volume and hazard score using the 2 factors incorporated so far, but I don't see a description of how the weighting will be accomplished for the other two factors, and these factors are not to be found in the algorithm as yet. The text suggests that the algorithm will instead be adjusted eventually to include them. The authors do give some very good ideas as to how the data will be assembled and analyzed to understand the influence of seasonality and production practices on Hazard, and hence sampling probability, but I can find nothing in the text describing how these factors will be weighted into the final sampling probability score.

The analysis of seasonality is well done, at least given the limited dataset available, and the Chi square analysis clearly shows a statistically significant difference in one of the clusters of months that should be reflected in the final methodology (although I am not convinced it will have a large influence on the final sampling frequencies given the importance placed in the algorithm on ensuring that all sites are sampled with reasonable frequency throughout the year). And the authors have identified what I believe to be the most significant practices that might influence hazard and have identified a good pool of institutions to participate in any subjective encoding needed to determine the influence of these practices on hazard. I simply can't comment on how this information will eventually appear in the algorithm because the studies have yet to be done.

I disagree that sample history is weighted "proportionate to the increased probability of detecting contaminated lots". I don't see a justification for claiming that a facility that has had contamination in the past several months (the criterion used) is 5 times as likely to show contamination in the near future as one with no contamination in that period (which receives a Hazard Score of 1 in the algorithm). The probability of detecting contaminated lots certainly is HIGHER, but I doubt it is 5 times higher. The analysis performed by the authors shows that the odds ratio is closer to between 2 and 3 than it is to 5. So I am not sure the term "proportionate" is correct here, unless the authors mean this in something less than a strictly mathematical sense.

Overall, though, the weighting is reasonable, even if it can't be related directly to any probabilities. I don't consider this lack of a fully probabilistic interpretation to be a weakness, because the goal was simply to improve the allocation of sampling in a way that is more likely to catch public health threats than is the current methodology. And this new methodology accomplishes that goal to a reasonable degree, as shown in Tables such as Table 19.

Question 6. *Comment on the performance measures described in the report. Do they provide an objective measure of the program's performance? Why or why not? If not, provide alternative performance measures. Please keep in mind that the performance measures must be quantitative, objective and based on data readily obtainable by FSIS.*

I assume these performance measures are:

- *How effectively testing resources are utilized to monitor the greatest percentage of beef products generated*. This is both a completely reasonable objective, being public health protective, and one that the proposed methodology achieves. The tests done on the sampling to date, as reviewed in the document, clearly indicate that the proposed sampling methodology captures a significantly higher percentage of beef products than was the case under a methodology where each facility was equally likely to be sampled.

- *Another metric is the ratio of the prevalence of positives in the risk-based sample pool to the prevalence of positives in an unweighted random pool.* Table 19 demonstrates that this goal is being achieved. This is a very good metric, and the proposal as to how it would be determined is sound. I share the authors concern, however, that the low occurrence rates of contamination will make the comparison statistically unreliable, at least until several years of data are collected. Still, the proposed method is conceptually sound. Of the two proposed methods, I prefer the second (the one with side-by-side comparisons) since it makes for the most complete assessment of the differences between the two approaches to selecting sampling sites.

- *If the program is verifying the safety of a (i) greater portion of product and (ii) the riskiest portion of product, then it is reducing the exposure of consumers to E. coli O157-contaminated ground beef (see above measures).* Again, Table 19 makes me comfortable that this goal is being achieved. This is another good measure and, as the authors state, this measure is assured by the first two measures above. The results presented in the document convince me *a priori* that the new methodology will sample a greater portion of the product. Whether it will also capture the riskiest portion depends on whether one believes that a past instance of contamination at a facility means that facility is also more likely than others (with no prior contamination in the past 4 months) to be contaminated again. The limited data in the document do suggest that this assumption is at least reasonable (an odds ratio of between 2 and 3), although it is not possible at present to estimate this conditional probability well and hence to estimate the improvement in the percentage of risky product identified in the old and new systems.

- *Another measure of public health impact can be estimated by predicting the number of human illnesses directly prevented by detection of an E. coli O157:H7-positive lot by the program*. This measure requires precisely the information I mentioned in the previous bullet, and has the same limitation in reliability. It also relies quite heavily on being able to track disease outbreaks to a specific exposure pathway such as beef, unless the authors mean to use an exposure-response relationship to estimate disease incidence rather than measuring incidence directly (as in the next bullet). The difficulty in the former approach is that it is necessary to not only predict exposure-response functions (which are not yet well developed in the literature), but to also model the intervening influence of food transport, storage, preparation and consumption. There are significant uncertainties associated with these steps, particularly due to the re-growth of the microbes as beef is stored. The approach is, however, feasible in principle. I and my colleagues developed such a modeling approach for eggs that would be analogous (see for example H. Latimer, L. Jaykus, R. Morales, P. Cowen and D. Crawford-Brown, "Sensitivity Analysis of Salmonella enteritidis Levels in Contaminated Shell Eggs using a Biphasic Growth Model", International Journal of Food Microbiology, 75, 71, 2002.)

- *Finally, FSIS can use surveillance data from the Centers for Disease Control and Prevention (CDC) to estimate the number of E. coli O157:H7 illnesses due to ground beef*. I am the most skeptical of this approach, although it is the one that would provide the most direct evidence. It is difficult to ascribe outbreaks to a single exposure pathway, and there is the large problem of under-reporting of disease (all one generally measures is disease that has gotten to the point of requiring medical intervention). I would, therefore, give this method of verification the lowest priority.

Reviewer 3:

Review Report on E. coli O157:H7 Risk-based Sampling

a. Is the overall approach for allocating samples through a probabilistic draw sound? (a random draw where the probability of selection is weighted by risk) If not, what problems exist and how should they be addressed?

The approach employed in the developed algorithm is statistically sound and is the right way to improve over the existing practice. If implemented correctly, this will be a cost-effective and more accurate alternative to the current practice in controlling and monitoring the E. coli O157 in the beef industry.

b. Evaluate algorithm source code and mathematics.

1) Are the techniques (mathematics and equations) appropriate? If not, provide alternatives.

Yes. I think the mathematics and equations are correct. I suggest evaluating other forms of weights composition, in addition to the multiplicative one proposed now, to see if more efficient or sensitive alternative ones exist (see my comments to c, d and f).

2) The reviewer should examine and verify that the data analysis and source code are accurate.

I can not assess whether the source code are accurate - I did not get the demonstrating program to work as I am not that familiar with the program.

c. Does the algorithm accomplish the three objectives (described above and in the report)? Please be specific. If so describe why and if not suggest how it could be altered to better achieve the described objectives.

The algorithm is in the right direction for accomplishing the three objectives. The weights outlined in the report, namely the volume*(risk score) or v*h, addressed only partially the aims outlined in Objective 1. If the algorithm is fully implemented, i.e. after the proper score for the establishment's intervention and testing programs are incorporated to the procedure, it will be a more complete/accurate algorithm. As it

currently stands, the weights in the algorithm are only determined by two factors, the production volume and recent test results. There are some important details left unspecified, e.g., how the season and establishment's prevention programs will be weighted in the algorithm and how sensitive/reasonable the new weights will be. Furthermore, the composition of weights may also need evaluated further, e.g., one may consider additive form of the composted weights or the logarithm transformed forms (see comments to b (1)).

Objective 2 is accomplished. Objective 3 is accomplished in both newly proposed and existing algorithms.

d. The algorithm uses four general areas to assign the overall risk of causing O157:H7 illness (production volume, sample history, season and establishment practices). Comment specifically on the use of each of these areas. Are there additional factors that should be considered?

The selection of the four areas to determine the relative weights in the proposed algorithm is appropriate. The new weights allow the production volume in an establishment to play a factor in their relative weights. Compared with the current system, which assign every establishment same weight, the proposed algorithm indeed increases the proportion of the large establishments to be tested, and this translated into a bigger portion of the total U.S. beef production to be tested more frequently than before.

The data presented suggests that if an establishment has been tested positive for E. coli O157 in the recent past, then it is more likely that this establishment will be tested positive within a short time window after the initial test (within 4 months). This treatment makes sense as the establishment may need some time to identify the sources for possible contamination and corrective action may take some time to be effective. Increase the relative weights to one within a 4-month window as suggested is certainly warranted. It may be worth considering in the future to adjust the length of this window, or using a step down weighting system in a wider time window.

It should be noted that while the report suggested the season and establishment's practice as factors for assigning the sampling weights, these are not currently incorporated into the algorithm and there are no specific formula details on how the weights will be determined. The intention of using these to further determine the final relative sampling weights is right but how the appropriate weights that combines these factors should be calculated is an important question and remain to be unclear. Some simulation studies may be needed in the future to help identifying the weights that will yield efficient and sensitive sampling designs.

I am not familiar enough with the industry to suggest any other factors that might be predictive to the occurrence of the positive testing results. However, if there is a database available on the establishments, more formal analysis than the one used in

the report may be can be carried to identify what are the predictive factors. If such analysis was done and some factors are identified, then maybe the predicted risk from the regression can be used as alternative sampling weights, with some adjustment of the production volume. One may compare the relative advantage of these different weighting systems.

e. *Each of the four areas (production volume, sample history, season and establishment practices) is assigned a weight in the algorithm to determine the overall probability of sampling for each establishment. In some cases weighting is proportionate (i.e. seasonality and sample history are weighted proportionate to the increased probability of detecting contaminated lots) while in other cases the weighting needs to be balanced with the overall objectives of the program (i.e. production volume weighting is more complex due to additional sampling concerns). Comment specifically on the method of weighting each area. Is the weighting consistent with the three objectives?*

First of all, there is no weight adjustment for the season and establishment practices implemented by this report, although the authors have discussed the empirical evidence and needs for considering them in the algorithm.

The proposed weights for the volume and risk score (last positive test, to be more precise) is consistent with the stated objectives. There is a need to evaluate the sensitivity and exploration of other forms and scales of the weights composition. For example, the current algorithm assigns a rather arbitrary multiplicative factor of 5 if an establishment has been tested for E coli. positive in the last 4 months, compared with a multiplicative factor of 1 for those tested negative. For a given set of constraints, one could in theory find out what is an optimal factor to achieve the most cost-effective design. For practical purpose, one could study the impact of the factors by comparing different choices of the values in simulation study.

f. *Comment on the performance measures described in the report. Do they provide an objective measure of the program's performance? Why or why not? If not, provide alternative performance measures. Please keep in mind that the performance measures must be quantitative, objective and based on data readily obtainable by FSIS.*

The performance measures for Objective 2 and 3 are objective and convincing. There is clearly an increased proportion of the establishments with high production volumes being tested under the proposed scheme (Table 6 – Table 10). However, quantitative measure of performance regarding Objective 1 is not demonstrated. The report suggested a good measure of effectiveness (4^{th} paragraph, p.30) but does not provide any results, demonstrating that the new scheme indeed catches more positive samples than the existing one.

This can be addressed with a simulation study where one mimics the parameters in the real situation and uses both algorithms to sample a fixed number of test samples.

By repeating this procedure for a large number of times independently, say 1000 times, one can look at the mean catch rate of the new sampling algorithm and compare it with the existing one. In this exercise, one can also evaluate difference weighting schemes mentioned earlier to see if one is more efficient than the other.

Reviewer 4:

Evaluation of risk-based sampling program for *Escherichia coli* O157:H7 in ground beef and beef trim

General remarks

The report "risk-based sampling for Escherichia coli O157:H7 in ground beef and beef trim" describes a comprehensive approach to improving the allocation of samples in the surveillance program by taking production volume and risk of contamination with *E. coli* O157:H7 into account. In the model currently available, only one risk factor (sample history) is included. In this review, I have tried to evaluate the final model, which is likely to improve the public health impact of the surveillance program considerably more than the simplified model available now. The report describes very well how additional risk factor information is going to be collected. Nevertheless, the challenging question on how to combine information on different risk factors is not addressed in the report. Also, the data presented in the report and the model only refer to ground beef. Some more information on beef trim production would be helpful to decide whether the results presented in the report also apply to beef trim producers.

Specific remarks to questions described in charge to peer reviewers

a. The approach for allocating samples through a probabilistic draw is sound and has a great potential to improve the efficiency of the surveillance program to protect consumers from *E. coli* O157:H7 infection. The surveillance program is very well suited for risk-based sampling. *E. coli* O157:H7 is present in ground beef and trim at a low prevalence. The risk of contamination is unevenly distributed among establishments, and extensive data on risk factors is available from the literature and past testing. In this situation, the many resources necessary for detecting contaminated lots through random sampling can be utilized more efficiently by targeting the samples to those establishments which have the greatest risk of contamination.

   If risk factor information is biased, risk-based sampling may perform worse than random sampling. For example, an establishment classified incorrectly as low production volume will have a small chance of selection for sampling, even though it poses a relatively high risk. The approach presented here addresses this potential problem very well, because establishments classified as low risk are also sampled at a reasonable frequency. This way, risk factor information can be regularly updated as new data on infection risk of different types of establishments become available.

b. The model is very user-friendly and well documented. The source code is flexible and structured in a way which allows modification of the algorithm also by less experienced visual basic programmers. In addition to the documentation provided, a

list of the variables used in the visual basic code and their meanings would be helpful. The techniques are appropriate and correctly implemented in the source code. Details on suggestions for alternative mathematical solutions for the algorithm are discussed under e. A minor drawback of the way the source code is implemented is that the model does not run under versions of Excel with a date format different from the US date format. This could be avoided by working with dates in a general number format rather than the format 'mmddyy'. I also encountered problems when opening the file from the list of recent files in Excel rather than opening it over the menu system. In this case, the visual basic program tries to save the new file under the default directory (which happens to be read-only on my computer) rather than under the directory where the file 'O157 sampling algorithm' is saved. I also encountered problems when selecting a date before 2005 for the sampling ('runtime error 6'). Unfortunately, I did not find out what caused this error.

c. All three stated objectives are accomplished by the algorithm. For objective two (verifying a greater portion of the supply), it is difficult to evaluate how much of the supply can be verified with the program. One sample per establishment will be less representative of the total daily production of a large establishment compared to establishments with a lower production. Objective three is relatively subjective, because there is no clear definition of what is meant by a reasonable frequency. In the suggested program, a median of 4 samples per year are collected from establishments in the lowest risk category. While this should be sufficient to ensure a reasonable frequency, a better definition of this objective would help to improve the algorithm. For example it could be stated that at least X% of the samples should be allocated randomly, at least n samples should be available for establishments in the lowest risk category, or the expected number of samplings should be at least X per year for all establishments.

d. In the current model, only production volume and sample history are used to determine the probability of selection. Because very few establishments have a history of positive test results, the probability of sampling in the current model is mainly determined by production volume. There is a very high correlation between production volume category and plant probability points (Spearman's rho for January 2006 data = 0.997). The current algorithm results in a random sample weighted by production volume, with re-testing of the few establishments with a history of contamination. This result could have been achieved with a much less sophisticated model. In order to utilize the potential of risk-based sampling, additional risk factors need to be taken into account. Of the factors discussed, establishment practices such as process control, interventions, employee training and testing programs are the most important ones to include. This helps to identify establishments with a decreased risk of contamination. In addition, recording these factors will provide an incentive to establishments to improve their manufacturing practice. If it is practical to vary sampling intensity by season, this risk factor should also be included as described in the report. Participation in quality assurance programs and geographic location of the establishment or the suppliers could be additional risk factors to consider. In the current program, two different sampling frames are planned for ground beef and beef

trim producers. They could be combined in one sampling frame, with type of producer taken into account as a risk factor (provided that the other risk factors are similar for both types of establishments).

e. The method to assign a weight to sample history is straightforward and well justified. An explanation could be added why the model assigns an increased hazard score for all establishments positive within 150 days before sampling, whereas the report describes an increased risk within 120 days after a positive sample. Strictly speaking, the relative risk or risk ratio (RR) would be a more appropriate measure for cross sectional data than the Odds ratio (OR). However, because E. coli O157:H7 contamination is a rare event, both measures are almost identical (OR=4.86; RR=4.79). The effect of season could be described accordingly. The RR for April-October compared to November-March would be 2.14 (95% CI 1.47-3.11), the OR 2.14 (95% CI 1.47-3.12). Different risk factors could be combined to a common hazard score by two different methods (1) multiplication of the RR or OR values for different risk factors (i.e. an establishment with a history of contamination in the season April-October would be assigned 10 hazard points, an establishment with no history of contamination 2 points); (2) calculation of OR in a multiple logistic regression model. The second method would be preferable, because the estimates for each risk factor could be corrected for the other risk factors. However, with this method risk factor information from scientific literature cannot be included. Estimating the effect of establishment practices will greatly rely on literature, because obtaining this information retrospectively may result in biased data.

   The way in which production volume is currently handled in the algorithm has several disadvantages. First, production volume has an effect on the exposure of consumers as well as on the risk of an establishment to produce contaminated meat. According to the data provided in the report, establishments in category 3 have a 2 times greater risk of positive samples compared to category 4 establishments (RR=2.34, OR=2.35). This information is currently not used in the algorithm. It could be argued that this risk factor information should be included in the hazard score rather than in the volume score, because the latter reflects exposure of consumers. Second, volume score is recorded on a flexible scale (1-4, 2-5 or 3-6) to fulfill the decision criteria stated in the report. If the higher scale is used, the relative weight of volume score compared to hazard score is greater than if the lower scale is used (production volume is more important in driving sampling probability if the higher scale is used). Some kind of standardization should be applied to avoid this problem (either choose a large enough scale to ensure that the criteria can be met with a fixed value for the highest category, or allow decimal numbers in the score). Finally, the importance of volume score for driving sampling probability depends on the scale of the hazard score, and on the distribution of risk factors among the establishments. In the current model, volume score is the primary driver of sampling frequency, even though it is stated in the report that this should be avoided. This is the case because the majority of establishments has a hazard score of 1. It would be beneficial to achieve more transparency on how the relative importance of volume score and hazard score is determined. One option would be to standardize the production

volume point values and the hazard point values to a scale with a common maximum and minimum value before multiplying them to obtain the plant probability points. Another (somewhat more complicated) option would be to define a fixed relative importance of volume score and hazard score in determining sampling probability. In this case, the volume score points would be corrected by multiplying its value for each establishment with a correction factor X, which depends on the relation of the sum of all volume points V_i to the sum of all hazard points H_i, where Y and Z describe the relative importance of volume score and hazard score in determining sampling probability (e.g. Y=40%, Z=60%):

$$X = \frac{Y \cdot \sum_{i=1}^{n} H_i}{Z \cdot \sum_{i=1}^{n} V_i}$$

f. Both measures suggested to evaluate how effectively resources are being used are valid and provide good information on different aspects of the performance of the program. They need to be balanced because maximizing one of these measures will result in a poorer performance of the other measure. Running a random and a risk-based sampling program in parallel would be a huge effort. An alternative could be a stochastic simulation which compares the effectiveness of both sampling strategies, while taking the uncertainty of risk factor and prevalence information into account. An additional useful measure could be the cost associated with detection of one contaminated sample. Measuring the public health impact of the sampling program will be quite difficult. The greatest public health impact of the improved surveillance program will probably be through prevention of contamination of ground beef and beef trim by sensitizing producers to the risk of *E. coli* O157:H7, and by providing incentives for good manufacturing practices. Thus, the measures suggested here are likely to underestimate the effect of the sampling program on public health.

Additional remarks

- Table 2 of the report has incorrect headings. Instead of "% Meat" the heading should read proportion (or the numbers changed to a % format). Also, you might want to explain that the proportion of trim is referenced towards the meat, not the total carcass weight.
- The excel file 'Testing data' contains 4 invalid dates in the column 'collect date' (1900 instead of 2000 and 1901 instead of 2001). This error should not have an effect on the analyses, though, because the column 'analysis end date' with corrected dates is used instead of 'collect date'.

Reviewer 5:

Summary evaluation

The proposed risk-based probabilistic sampling algorithm for E.coli O157:H7 represents another positive step towards FSIS's continuing development of a risk-based approach to meat inspection and food safety. The document was clearly written and structured, the basis for the risk calculations were well described, and the visual Basic code was well documented. Some minor inconsistencies and omissions were present in the report but, in my opinion, these do not adversely impact its utility. Suggested modifications are documented in the body of the report. One important assumption relative to data comparability, i.e. that sampling and testing protocols are not confounded with establishment (and production values) was not mentioned. Second, an implicit assumption is made that prevalence is the correct measure although concentration of E.coli O157 is obviously more relevant. Third, once the algorithm is implemented, the sampling protocol should be modified on no more than an annual basis to reflect changes in risk from modification in slaughterhouse management practices and other risk factors. Measurement of the impact of new system on public health impact will be challenging.

Specific tasks

g. Is the overall approach for allocating samples through a probabilistic draw sound? (a random draw where the probability of selection is weighted by risk) If not, what problems exist and how should they be addressed?

Yes, the approach is sound and well justified based on known risk factors. As information on other risk factors becomes available (and is routinely) measurable, then probabilities used in the algorithm can be updated. A system needs to be implemented to ensure that the algorithm used in a specific year incorporates the most recent/relevant information. I recommend that FSIS develops a plan and a data analysis strategy to do this to minimize possible criticism that the sampling probabilities are outdated.

h. Evaluate algorithm source code and mathematics.

1) **Are the techniques (mathematics and equations) appropriate? If not, provide alternatives.**
2) **The reviewer should examine and verify that the data analysis and source code are accurate.**

Visual basic is not an area with which I am very familiar so my comments are limited. On page 38, why give a value of 0 if there are no data available? What are

the reasons for missing data – no recording of information, stopped production? If a small establishment is not sampled once in the first 11 months, will its sampling probability in the twelfth month be 1?

The odds ratio (OR) calculations (Tables 9 and 10) are not exactly correct because a comparison group of all samples is used. Standard epidemiologic practice dictates that the 2 exposure categories should be mutually exclusive, e.g. factor positive versus factor negative if the exposure is binary. However, given the rareness of the outcome, a positive O157 result, they do not change dramatically.

Also, on pages 20 to 22, I strongly suggest that the seasonal data (esp. data in table 14) be presented as OR to be consistent with presentation of data for sample testing history. Based on my calculations, the OR is 2.15 for a comparison of the two seasonal categories. One final point about the OR calculations -- the confidence intervals do not account for the clustered sampling design and hence, are narrower than they should be.

i. Does the algorithm accomplish the three objectives (described above and in the report)? Please be specific. If so describe why and if not suggest how it could be altered to better achieve the described objectives.

The stated objectives of the algorithm are:

- To increase the proportion of FSIS samples taken at establishments that are more likely to produce product contaminated with *E. coli* O157:H7.

This is achieved by increasing sampling in establishments with a sample history and high production output.

- To allocate FSIS resources more efficiently by verifying a greater portion of the U.S. trim and ground beef supply with the same number of samples as the current program.

This assumes that the risk-based estimates are unbiased and not confounded by other unknown factors. The greater the magnitude of the risk factor(s), the more beneficial the use of a risk-based approach to sampling.

- To verify *all* eligible establishments at a reasonable frequency regardless of an establishment's production volume, interventions, or predicted public health risk associated with their product.

This is achieved by ensuring that all establishments are sampled. "Reasonable frequency" is not defined but might be as low as a single sample in some establishments. The changes in the program sampling are clearly shown in

Figures 6 to 10. Because intervention data have not been analyzed, it is not possible to comment on this aspect.

j. The algorithm uses four general areas to assign the overall risk of causing O157:H7 illness (production volume, sample history, season and establishment practices). Comment specifically on the use of each of these areas. Are there additional factors that should be considered?

First, the use of production volume is well justified since large establishments produce the greatest volume of servings and theoretically, more contaminated servings given a constant individual risk of a contaminated serving. Second, the choice of sample history was well justified based on the odds ratio estimate of 4.86. However as the authors indicate on page 18, collection of additional follow-up samples over time will greatly improve the ability to draw conclusions about the importance of sample history *E. coli* O157:H7 risk.

Third, the seasonal risk of human illness associated with consumption of ground beef products closely parallels the seasonal risk of *E. coli* O157:H7 positive samples and hence, there is sound justification for inclusion of season in the sampling design. Whether or not laboratories can handle the increased capacity at that time of year is a logistical issue that will need clarification. Finally, establishment practices that reduce the prevalence of O157 will be included. Since these are not available for evaluation, then I can only indicate that philosophically this is a good choice.

k. Each of the four areas (production volume, sample history, season and establishment practices) is assigned a weight in the algorithm to determine the overall probability of sampling for each establishment. In some cases weighting is proportionate (i.e. seasonality and sample history are weighted proportionate to the increased probability of detecting contaminated lots) while in other cases the weighting needs to be balanced with the overall objectives of the program (i.e. production volume weighting is more complex due to additional sampling concerns). Comment specifically on the method of weighting each area. Is the weighting consistent with the three objectives?

The weighting of seasonality and sampling history is based on individual odds ratio estimates and is consistent with the stated objectives. However, it is possible that the individual OR estimates could be confounded and hence, it might be prudent to do a stratified analysis of sample history by season to ensure no confounding (or effect modification).

l. Comment on the performance measures described in the report. Do they provide an objective measure of the program's performance? Why or why not? If not, provide alternative performance measures. Please keep in mind that the performance measures must be quantitative, objective and based on data readily obtainable by FSIS.

Two categories of performance measures are proposed. Both are reasonable albeit with some limitations. I am unable to suggest better alternatives but rather make some comments.

The *first* measure is the effectiveness of resource allocation to verify the safety of the trim and raw ground beef supply. Two alternatives are proposed. First, the percentage of supply verified each month will be estimated and second, the portion of potentially contaminated product that is verified will also be estimated. As the authors indicate, a random, unweighted sampling program needs to be run side-by-side with the algorithm and results of the 2 programs directly compared. Decisions relative to whether this is a viable choice would need to be based on cost and likely sample sizes to detect meaningful differences in prevalence. To a large extent, the ability to make sound inferences about the effectiveness of the system will depend on the availability of these data.

The *second* measure, the public health impact of the sampling program, is much more difficult to quantify because of the number of intervening factors between reduced product contamination and a lower burden of human illness. The goal is laudable and one with which we collectively struggle. The authors justify the program as reducing the exposure of consumers to *E. coli* O157-contaminated ground beef. Theoretically this should be true, but without a side-by-side evaluation of the risk-based and the traditional program, this contention it is purely speculative. The second choice (prediction of the number of human illnesses directly prevented by detection of an *E. coli* O157:H7-positive lot by the program) is vague. Does the current risk model for O157:H7 allow this to be done readily? The use of CDC surveillance data is reasonable but a limitation now especially for sporadic *E.coli* illness is that consumers have become more aware of alternate sources of *E.coli* contamination and might be less likely to report an exposure to ground beef.

Additional comments related to report presentation and clarity (P= page, L = line)

P4 – the term “poor performers” is used here and again on page 34 – for clarity, a definition (or an example) of a ‘poor performing establishment” should really be given in the document.

P5 – hyphenate “high-risk” when used as an adjective – same comment applies on page 22.

P5 – last sentence – suggest change to “… should therefore increase the public health impact through more efficient allocation of FSIS resources.”

P6 – superscript for reference 5 to 9 is missing - were these references used?

P7 – Isn’t it true that other sources, e.g. produce, are currently as least as important and under increased surveillance for contamination? It would helpful to incorporate food attribution data, but I assume these are not available.

P7 – references 20 to 23 are not in the reference list.

P8 – why was the last 4 months chosen as the time window for sample history?

P9- define "high prevalence season" here, or at least refer the reader to section where the seasonal data are evaluated (pages 19 to 22).

P9 - The sentence "The risk-based sampling program will be more representative of the beef supply than the current sampling program that provides random sampling by establishment (i.e., without consideration of the amount of product produced, interventions, sampling history, etc.)" is not really correct. Only the first point "without consideration of the amount of product produced" is relevant to beef supply per se.

P9 and throughout the document – what was the justification for not using the most recent data, i.e. 2006, in the analyses?

P11 – why is there a need to assume an "upper bound of 500,000 lbs" for category 1? Aren't there data relative to production levels for these establishments?

P11, Table 2 – if the column headings (% meat and % trim) are correct, then the respective values should be 70,70,70,70 and 18,18,53,90.

P11 – define "MT03" samples here since the acronym is used here for the first time in the document. This could be readily achieved by copying or moving the relevant definition which appears in the last paragraph on page 14.

P12, Table 4 – different numbers of establishments are used in this table compared with other places in the document -- here 1668 versus 1536 (previously) and 1540 (later). Perhaps a single sentence could be added somewhere to account for the differences.

P12, Table 5 – change "no cat" to "not available" or "N/A" to be consistent with usage on page 11.

P12, Tables 4 and 5 - remove % in 2 right-most columns because this is in the table column headers already (ie., % total production volume, % total samples).

P17 – wording of first sentence below the table should be change to reflect the comparison group. New wording suggestion – "Thus, random samples collected within 120 days of a previous positive sample at the same establishment are more likely to be positive for *E. coli* O157:H7 than samples collected within 120 days of a previous negative sample" . This assumes that the comparison categories will be mutually exclusive.

P14 to 18 – It would be helpful to have some background information on sampling methods, numbers of laboratories doing the testing and variability, if any, in *E.coli* test

protocols. A question that might arise is whether some establishments have more positive results just because their sampling and testing protocols are more rigorous and sensitive than protocols used at other establishments.

P23, L2 – should be “prevalence”, not “incidence” since it is snapshot picture in time.

P24, Table 16 – “does test” not “does tests”.

P25, L4-5 – this sentence implies that the survey has been done. Is this correct?

P26, second point – the reality of slaughter plants is that there are often a series of sequential mitigations to reduce the frequency of *E.coli*. How will the joint effect of mitigations be considered in this evaluation?

P27, last 2 lines – A brief explanation (or a cross-referencing to Table 9 on page 16) should be given to justify the choice of hazard scores.

P28, Figure 2 – why do the production volume box and the O157 result box both have the same designator, M2K?

P29, last line – should be “data are”.

P30 – I question the correctness of the statement “The overarching goal of FSIS’s *E. coli* O157:H7 testing program is to help ensure that industry is producing raw ground beef that is free from O157:H7 contamination”. Shouldn’t it be minimal risk or a similar modification of the wording? Given the small sample that is collected and tested, and the imperfect sensitivity of testing methods, the statement as written is misleading.

P31, Table 19 heading – hyphenate “risk-based”.

P31, Table 19 footnote – “unweighted” not “unwweighted”. Also last sentence seems incomplete. Perhaps it should read “ In the proposed risk-based weighted sampling program, an estimated 8,000 annual samples…..”

P34 – suggest deletion of sentence “Establishments with recent *E. coli* O157:H7-positives will be sampled monthly for four months” since this is at odds with the last sentence in the same paragraph and with statements made earlier in the document about no establishment being sampled with probability of 0 or 1.

P34 – why 1540 ground samples when 1536 were presented in Table 1?

P35 – how will a “good testing program” be defined?