J Vet Diagn Invest 12:118–125 (2000)

A statistical model for assessing sample size for bacterial colony selection: a case study of *Escherichia coli* and avian cellulitis

Randall S. Singer, Wesley O. Johnson, Joan S. Jeffrey, Richard P. Chin, Tim E. Carpenter, E. Rob Atwill, Dwight C. Hirsh

Abstract. A general problem for microbiologists is determining the number of phenotypically similar colonies growing on an agar plate that must be analyzed in order to be confident of identifying all of the different strains present in the sample. If a specified number of colonies is picked from a plate on which the number of unique strains of bacteria is unknown, assigning a probability of correctly identifying all of the strains present on the plate is not a simple task. With *Escherichia coli* of avian cellulitis origin as a case study, a statistical model was designed that would delineate sample sizes for efficient and consistent identification of all the strains of phenotypically similar bacteria in a clinical sample. This model enables the microbiologist to calculate the probability that all of the strains contained within the sample are correctly identified and to generate probability-based sample sizes for colony identification. The probability of cellulitis lesions containing a single strain of *E. coli* was 95.4%. If one *E. coli* strain is observed out of three colonies randomly selected from a future agar plate, the probability is 98.8% that only one strain is on the plate. These results are specific for this cellulitis *E. coli* scenario. For systems in which the number of bacterial strains per sample is variable, this model provides a quantitative means by which sample sizes can be determined.

Many samples that are cultured for bacterial pathogens are expected to contain a genetically heterogeneous population of bacteria. Often, the phenotypic appearance of the bacterial colonies will not reflect the genetic variability among different strains in the sample. This strain determination is made by additional diagnostics, such as serogrouping, antibiotic resistance profiles, plasmid profiles, or DNA fingerprinting methods. A general problem for microbiologists is determining the number of phenotypically similar colonies growing on an agar plate that must be analyzed in order to identify all of the different strains present in the sample. If the goal of sampling is to identify the number of different strains present in each sample, colony morphology will often fail to provide the answer. Without the knowledge of the total number of different strains on the plate, assigning a probability of correctly identifying all of the strains present on the plate for a given number of identified colonies is extremely difficult. If samples typically contain a single strain, then the identification of more than one colony is a waste of time and money. However, if samples typically contain more than one strain, then the selection of a single colony per plate would lead to biased inferences.

This problem is relevant to studies of avian cellulitis in broiler chickens. This condition is characterized by a diffuse inflammatory reaction secondary to a subcutaneous infection.[4,8] Although avian cellulitis in broilers is a multifactorial process,[2,3] numerous investigators have experimentally linked the presence of *Escherichia coli* with the condition.[11,12] Because this lesion is likely initiated by a breach in the broiler integument followed by infection with environmental *E. coli*, the possibility exists that the lesions contain heterogeneous populations of *E. coli.* If there are multiple strains of *E. coli* per lesion, more colonies per clinical sample will have to be analyzed or the sample size of the number of lesions sampled will have to be increased in order to accurately assess the distribution of cellulitis-associated *E. coli* among broiler houses.

DNA fingerprinting techniques are being used to assess the diversity, distribution, and persistence of cellulitis-associated *E. coli.* The ability to detect the number of different DNA fingerprints of *E. coli* residing within a lesion is of primary importance in order to track pathogenic *E. coli* and to draw accurate inferences about the environmental distribution of this organism. In addition, sample size calculations for the number of lesions to be sampled and the number of bacterial colonies to be selected from each agar plate will depend on the number of different DNA fingerprints that are growing within the lesion. Because

From the Departments of Medicine and Epidemiology (Singer, Carpenter), Population Health and Reproduction (Jeffrey, Atwill), Pathology, Microbiology, and Immunology (Hirsh), School of Veterinary Medicine and the Division of Statistics, University of California, Davis, CA 95616 (Johnson), and the California Veterinary Diagnostic Laboratory System, Fresno Branch, School of Veterinary Medicine, University of California, Davis, CA 95616 (Chin). Current address (Singer): the Department of Veterinary Pathobiology, University of Illinois, 2001 South Lincoln Avenue, Urbana, IL 61802.

Received for publication February 23, 1999.

MacConkey agar is commonly used to culture *E. coli* from each lesion, the majority of *E. coli* isolates from these lesions are lactose fermenting and appear as pink colonies. However, the presence of pink colonies does not indicate the number of different DNA fingerprints in the sample. Colony morphology cannot be used to determine the number of different strains (DNA fingerprints) in the sample.

The primary reason that determining the number of colonies to be analyzed per plate is difficult is that the actual number and distribution of strains in the sample is unknown. If the number and distribution of bacterial strains are known, the probability of identifying all strains can be calculated. For example, with two strains growing in equal numbers on an agar plate, the probability of identifying both strains, assuming three colonies are randomly selected, is calculated[6] as:

$$\sum_{x=1}^{n-1} \binom{n}{x} (p)^x (1-p)^{n-x} \equiv \sum_{x=1}^{2} \binom{3}{x} (0.5)^x (0.5)^{3-x} = 0.75;$$

n is the number of colonies selected, x is the number of colonies of one strain selected from the plate, and p is the probability of selecting a colony of that strain. It must be emphasized that the formula for the binomial distribution shown above assumes that the two strains are in equal concentration. With a greater number of strains, a multinomial distribution would be used to make the calculation. The generalization of the multinomial distribution is shown in the Appendix. Without knowledge of the actual number of strains on the plate, solving the problem is more difficult. This situation requires the imputation of estimates for the missing data of the true number of strains on the plate.

The objectives of this study were to 1) design a statistical model that would delineate sample sizes for efficient and consistent identification of all the strains of phenotypically similar bacteria in a clinical sample, 2) create a model that will enable a microbiologist to calculate the probability that all of the strains contained within the sample are identified, and 3) apply this model to the research of *E. coli* and avian cellulitis in broilers in order to determine the number of *E. coli* colonies that need to be analyzed in future studies.

Materials and methods

Sampling of E. coli from broilers. Twenty-four broilers with cellulitis lesions were necropsied. The birds originated from a single integrated broiler company in California. Each cellulitis lesion was cultured onto MacConkey and blood agar plates. Initially, the assumption was made that there would be three or fewer strains per lesion. This assumption was based on the opinion of an expert microbiologist as well as the observation that the majority of the lactose-fermenting colonies from cellulitis lesions had similar morphologies. If a high proportion of the lesions were heterogeneous, then this estimate would be increased. Three individually isolated lactose-positive colonies were randomly selected from the MacConkey plates and identified as *E. coli* by standard biochemical techniques.[7] Selecting three colonies for developing the initial model would result in a 75% probability of detecting two strains if there really were two strains on the plate (equation shown above). Although this sample size offered only a 22% chance of detecting three strains if there really were three, this uncertainty is accounted for in the model. In addition, the presence of three strains in a lesion was expected to be a rare occurrence.

Agar plate and statistical model assumptions. Several assumptions are typically made when agar plates are used in clinical diagnostic situations. First, it is often assumed that the strains that grow on the agar plates are representative of the strains contained within the sample. This assumption was not tested in this study. Second, there is often the implicit assumption that there are equal concentrations of strains within the sample and therefore on the agar plates. Because colonies are often picked from isolated areas of the plate, those strains that are in highest concentration are selected. Consequently, by drawing inferences about the strains present in a lesion when picking colonies from this distant portion of the plate, the assumption is made that all representative strains are in the high concentration portion of the plate. Therefore, the implicit assumption is made that all strains are in an approximate equal concentration. Although this assumption was not tested, the model was developed to allow for the scenario in which strains are not present in equal concentrations. Finally, it is often assumed either that different strains will have equal growth rates on the agar plate or that different strains will not compete with each other on a plate. This assumption was tested for *E. coli* of avian cellulitis in the following experiment.

The relationship between the ratio of the concentrations of different strains that were inoculated onto a plate and the relative proportion of individual colonies of each strain that subsequently grew on the plate was studied. Mutants of wild-type *E. coli* isolates (WT) originally isolated from avian cellulitis lesions were created that were resistant to 100 μg/ml nalidixic acid[a] (NAL) or to 100 μg/ml rifampicin[a] (RIF). All three strains were then grown overnight in brain–heart infusion (BHI) broth such that the final concentrations of WT, NAL, and RIF were 4.0×10^8 colony-forming units (cfu)/ml, 4.1×10^8 cfu/ml, and 4.4×10^8 cfu/ml, respectively. The strains were combined in various concentrations, and the ratios of each strain (WT : NAL : RIF) were 1:1:0, 2:1:0, 5:1:0, 1:1:1, 2:1:1, 2:2:1, 5:1:1, 5:2:1, and 5:5:1. Each combination was then streaked onto two different BHI agar plates and grown overnight. All individually isolated colonies from each plate were then transferred to three different BHI agar plates, one containing NAL (100 μg/ml), one containing RIF (100 μg/ml), and one pure BHI plate. This procedure enabled the identification of the strain of each colony on the original BHI agar plates. The number of individually isolated colonies transferred for each replicate varied from 8 to 18 colonies.

The observed numbers of colonies of each strain that grew on the plates were compared with the numbers expected on the basis of the initial concentration of the strain in the inoculum. For each inoculum, there were two plates for which

the number of colonies of each strain was observed. Expected numbers for each strain were calculated according to the original concentration ratios of the inoculum. For example, if 12 colonies formed and the concentration ratio in the inoculum was 1:1:1, the expected counts would be four, four, and four. For each plate, a Pearson χ^2 statistic was calculated, with degrees of freedom under the null hypothesis equal to the number of strains in the inoculum minus 1. The independent χ^2 statistics were pooled over all plates to obtain an overall χ^2 statistic with degrees of freedom equal to the sum of the degrees of freedom for each individual plate. Thus, this overall χ^2 statistic was comprised of the results of 18 different agar plates. Because the smallest expected value was greater than 1, the χ^2 assumption was justified.[1] The assumption of equal growth rates on the agar plate was rejected if $P < 0.05$ for the pooled χ^2 statistic. In order to determine if a rare strain was negatively selected for growth on the agar plate, those scenarios in which at least one strain in the inoculum was in concentration $\leq 1/6$ of the other strains (5:1:0, 5:1:1, 5:2:1, and 5:5:1) were analyzed separately. An overall χ^2 statistic was calculated as described previously. This overall χ^2 statistic was comprised of the results of eight different agar plates.

Finally, a sampling distribution was required for the process of picking colonies from agar plates. The model calculates the probability that a certain number of strains would be observed in the colonies analyzed, conditional on the number of strains that were actually on the plate. It was assumed that the counts for the number of observed colonies of each strain had a multinomial distribution (binomial in the case of two strains). See Appendix for details.

Pulsed-field gel electrophoresis (PFGE) of E. coli from broilers. DNA fingerprinting was performed on all isolates by PFGE. For extraction of genomic DNA, the CHEF Bacterial Genomic DNA Plug Kit was utilized as per manufacturer's instructions.[b] The agarose plugs were digested with 20 U of restriction endonucleases *Not*I and *Xba*I[c] in separate digestions at 37 C overnight. PFGE was performed with a 1.2% agarose gel on a CHEF III apparatus[b] in 0.5× Tris-borate–ethylenediaminetetraacetic acid (EDTA) buffer (45 mM Tris, 45 mM boric acid, 1 mM EDTA, pH 8.3) at 14 C and 200 V. Linearly ramped switching times of 5–50 sec and 1–40 sec were used for *Not*I and *Xba*I, respectively, and were applied over 22 hr. After PFGE, the gel was stained with ethidium bromide (0.2 μg/ml) and photographed under ultraviolet transillumination. Only PFGE fragments larger than 100 kb were considered in order to eliminate the potential influence of large plasmids. Although guidelines exist for delineating the epidemiologic relatedness of PFGE fingerprints,[16,17] for the purpose of this study, colonies with any differences in banding patterns were considered to be different strains.

Model development. A model is proposed to develop probability-based sample sizes for bacterial colony selection. The mode of inference is Bayesian.[5] This approach combines data that have been or will be observed with scientific information that is regarded separately from the data. The scientific input is given in the form of ''prior'' probability specifications, and a sampling distribution is generated for the observed data. Actual inferences are based on probability statements that are made conditional on the observed data and are consistent with the laws of probability. The object of this statistical inference is called a posterior distribution, or a collection of posterior probability statements, which are calculated after the data are observed.

For this model of *E. coli* of avian cellulitis, prior beliefs about the probability of one strain per lesion, two strains per lesion, or three strains per lesion were required. The assumption was made that the probability of a mixed infection would be less than the probability of a homogeneous population of *E. coli.* Therefore, the best prior probability estimates were 85%, 10%, and 5% for the probabilities of one, two, or three strains per lesion, respectively. The model was also evaluated with a range of other probability estimates, including symmetric prior information, which assigned 33.33% for all three prior probabilities. The testing of various probability estimates allowed the assessment of the sensitivity of the model to the selection of prior probabilities.

The Bayesian analysis was performed with statistical modeling software.[d] The method of implementing the analysis was iterative in nature and involved the simulation of random variates from distributions described in the Appendix.[5] A Gibbs' sampling routine[5,15] was used to estimate the unknown parameters of the model. The model used 5,000 iterations in order to achieve a high level of precision. Details of the model are provided in the Appendix.

The model was used to estimate the probability that a lesion would contain one, two, or three strains given the observed data and the prior probabilities. These probabilities can be considered as prevalences of lesions containing one, two, or three strains of *E. coli.* Bayesian intervals for these probabilities were also generated.

With Bayes' theorem,[5] the probability was calculated of there being a certain number of strains in a lesion, given the number of strains that were observed (see Appendix). For example, it was now possible to estimate the probability that there truly was one strain on a plate given that one strain was observed. Bayesian intervals for these probabilities were also generated.

Finally, the probability that the actual number of strains on the plate would be correctly identified was calculated (see Appendix). In these calculations, the number of colonies selected was varied in order to determine the optimal sample size for selecting bacterial colonies from agar plates.

Results

Validation of agar plate and statistical model assumptions. In the experiment in which *E. coli* mutants were created, the assumption of equal growth rates of different strains was tested. The observed number of colonies of each strain that grew was compared with the number expected on the basis of the ratio of each strain in the initial inoculum. The number of individually isolated colonies that grew on each plate varied from 8 to 18. When all plates were considered, the strains grew in proportion to their relative concentrations in the inoculum ($P > 0.995$). The analysis was also separated into those scenarios where one strain in

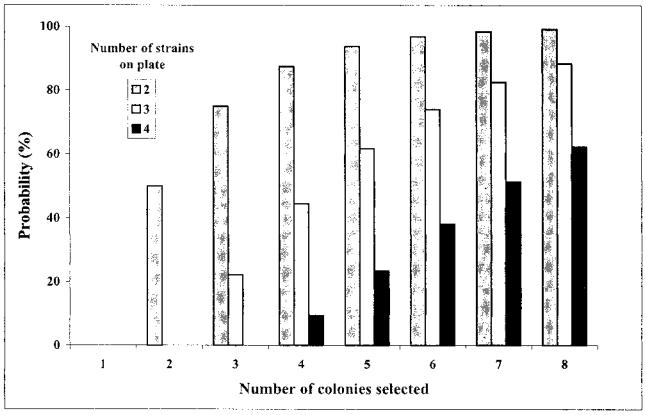


Figure 1. Probabilities of selecting all strains on an agar plate for various sample sizes (number of colonies selected) when the actual number of strains on the plate is known. These calculations assume a multinomial sampling distribution and an equal probability of selecting each strain.

Table 1. Median values and 90% Bayesian intervals (BI) for the parameters and conditional probabilities calculated in the model. The results are specific to this *Escherichia coli* scenario.

Parameter*	Median†	90% BI†
Θ_1	95.39	82.65, 99.52
Θ_2	4.46	0.46, 16.97
Θ_3	<0.01	<0.01, 0.96
P(i=1| j=1)	98.83	95.10, 99.88
P(i=2| j=1)	1.15	0.12, 4.87
P(i=3| j=1)	<0.01	<0.01, 0.12
P(i=2| j=2)	100.00	80.77, 100.00
P(i=3| j=2)	<0.01	<0.01, 19.23

* Θ_i represents the probability of a plate having "i" strains. P(i\| j) represents the conditional probability of there actually being "i" strains on the plate given that "j" strains were observed. This probability is calculated with Bayes' theorem.

† Median and the lower and upper 90% BI values were calculated as the 50%, 5%, and 95% quantiles of the corresponding simulated values for each parameter.

the inoculum was in concentration $\leq 1/6$ of the other strains in order to determine if a rare strain was negatively selected for growth on the agar plate, thus impairing the ability to detect rare strains. There was still no significant difference between the observed and expected number of colonies on the plates ($P = 0.96$).

The probability of correctly identifying all of the strains on a plate for various sample sizes (number of colonies selected) was calculated when the actual number of strains on the plate is known. These calculations assumed a multinomial distribution. This exercise was not specific to the *E. coli* example of this study but rather was based completely on the expected result of the probability distribution. A graphic representation of these probabilities with the multinomial distribution is shown in Fig. 1.

Sampling of E. coli from broilers. Of the 24 lesions sampled, 23 had one PFGE fingerprint type in the three colonies analyzed. All fingerprint bands of the three colonies from a plate were identical. One lesion had two different fingerprints, and these fingerprints differed in eight of the possible 20 bands.

Model results. Estimates for the parameters of interest were calculated with the 5,000 Gibbs' sampler iterations. The median of the simulated values for a given parameter was used as a point estimate for that parameter. The 90% Bayesian intervals (BI) were also determined for each parameter on the basis of the upper and lower 5% quantiles of the corresponding simulated values. Table 1 shows the median values and 90% BI for the probabilities of a lesion containing one, two, or three strains. In addition, Table 1 also shows the conditional probability estimates for the probability of a certain number of strains given the number of strains that were actually observed. It should be emphasized that the results shown in Table 1 are specific to this cellulitis *E. coli* scenario.

The probability of correctly observing all of the strains actually present on the plate was calculated for the scenario in which the actual number of strains on the plate is unknown. Figure 2 depicts the median probability of correctly identifying all of the strains on the plate for various sample sizes (number of colonies selected). In addition, 90% BI are provided for each probability. It again should be emphasized that the results shown in Fig. 2 are specific to this cellulitis *E. coli* scenario.

Finally, a sensitivity analysis was performed to determine the degree of influence that the selection of prior probabilities had on the posterior probabilities estimated by the model. This was accomplished by obtaining results with a range of probability estimates, including the aforementioned symmetric prior. The

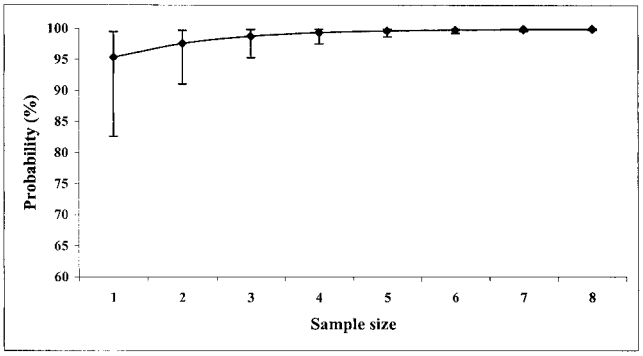


Figure 2. Median posterior probabilities of correctly identifying all of the *Escherichia coli* strains actually present on the plate for various sample sizes (number of colonies selected) when the actual number of strains on the plate is unknown. Error bars represent 90% Bayesian intervals for each probability. The results expressed in this figure are specific to the *E. coli* data presented.

median values for the probability of lesions containing one, two, or three strains when the symmetric prior was used were 93.19%, 2.82%, and 1.89%, respectively. Because these are medians of the 5,000 simulated values, they do not necessarily sum to unity. These estimates can be compared with the values in Table 1, which use the expert prior probability information as described previously. Although the expert prior probabilities are having an effect, the results from the two models are still quite similar.

Discussion

This model was designed so that researchers and diagnosticians can assign probabilities to bacteria colony selection from agar plates. Standard methods for calculating sample sizes are confounded in this case by the fact that the actual number of strains on a plate is unknown. In addition, the prevalence of lesions containing a certain number of strains is also unknown. This type of problem is perfectly suited to the Bayesian model and simulation method that has been described. The use of a Gibbs' sampling routine has provided a way to quantify the uncertainty associated with selecting bacterial colonies from agar plates.

In the study of *E. coli* and avian cellulitis, it was observed that, out of the three *E. coli* colonies that were randomly selected from each lesion, 23 of 24 lesions had a single strain. Thus, cellulitis lesions in broilers appear to be comprised of homogeneous populations of *E. coli.* The model then provided estimates of the prevalence of lesions containing one, two, or three strains. These estimates account for the possibility that, by selecting only three colonies per plate, all strains on the plate may not have been identified. Because of the fact that two strains were identified in one lesion, the possibility exists that multiple *E. coli* strains exist within a single lesion. Because other studies have found multiple different *E. coli* associated with cellulitis at the flock level,[10,13,14] the presence of multiple strains of *E. coli* within a lesion is not surprising.

The conditional probabilities shown in Table 1 were imputed from the prevalence estimates obtained at each iteration of the model. These conditional probabilities are useful in that they can be considered predictive probabilities for future plates. For example, if one *E. coli* strain is observed out of three colonies randomly selected from a future agar plate, there is a 98.83% probability that only one strain is on the plate.

The greatest benefit of this model is the ability to generate sample size calculations of the number of colonies to be selected from agar plates. These sample sizes are based on probabilities and not on historical records, tradition, or speculation. Cost and time efficiency can be weighed alongside the probability of success when determining the number of colonies to be picked. On the basis of the model, the selection of a single *E. coli* colony from an avian cellulitis lesion offers >95% assurance of correctly identifying the number of strains on the plate. This is because the majority of lesions contained a single strain. Although the lower bound of the 90% BI extends to 83%, the selection of a single colony is reasonable, especially in situations where multiple lesions are being assessed. However, it must be clearly understood that these calculations are based on avian cellulitis lesions in broilers in California. The probabilities expressed in this study should not be used for other bacterial systems. Data and prior information specific to the organism would have to be simulated in the model in order to obtain usable sample size and probability estimates.

In a Bayesian analysis, prior information concerning the unknown parameters is used to augment the observed data. This prior information is based on expert opinion and serves to initiate the estimation process of the model. Because the sample size in this study was 24 lesions and a prior sample size of one was used, the observed data and prior information accounted for 96% (24/25) and 4% (1/25) of the data in the model, respectively. A sensitivity analysis with a range of prior information, including the symmetric prior, was performed to ensure that the expert prior information did not exert an overly large influence on the model.[5] The parameter estimates obtained through the use of the expert prior information and the use of the symmetric prior information did not differ substantially. The probability of one strain per lesion decreased from 95.4% to 93.2%. The major difference was the increased probability of three strains per lesion when the symmetric prior was used. Because three strains per lesion were never observed, the model was not overly influenced by the choice of prior. On the contrary, the expert prior may have been more appropriate than the symmetric prior information.

Because the majority of lesions were homogeneous with respect to *E. coli*, the high probability of identifying all strains present by selecting a single colony is intuitive. The model enabled the assignment of a more accurate probability to this occurrence and to generate probability-based sample sizes for colony selection. The model would be even more beneficial in a system where a sample is expected to have variable numbers of strains. In studies of *Salmonella* spp. and *Campylobacter* spp. with fecal specimens and drag swabs, there is evidence of specimens having several strains per sample. A study of *Salmonella enteriditis* in humans[9] concluded that multiple colonies from a fecal sample should be identified in order to make accurate epidemiologic inferences about *S. enteriditis.* In many epidemiologic studies of *Salmonella* spp., XLT4 agar is used. Many of the clinically important *Salmonella* se-

rogroups will appear as black colonies on this agar because of the production of H_2S. Although the presence of black colonies suggests the presence of *Salmonella* spp., it does not indicate the number of different *Salmonella* serogroups present. This statistical model could then be used to determine the expected number of *Salmonella* serogroups in the sample as well as the number of black colonies that should be identified from the XLT4 agar. Without this model, the microbiologist has no way to ascertain the probability of correctly identifying all of the *Salmonella* serogroups that were actually contained in the sample.

A Bayesian model such as the one described here has great utility for a diagnostic laboratory or for researchers who are conducting long-term projects. After the initial model is run and sample size calculations are performed, all new future data that are collected can then be incorporated into the model to produce continually updated probability estimates. This type of adaptive model lends itself well to changes in the prevalence of an organism over time. For systems in which the number of bacterial strains per sample is variable, this model provides a quantitative means by which sample sizes can be determined as well as a method for estimating the probability that an observed result represents the actual situation.

Acknowledgements

This work was supported by a Division of Agricultural and Natural Resources competitive grant and by the Center for Food Animal Health, University of California–Davis. We thank Cara Cooke, Lori Hansen, Michelle Ganci, and Vera Perez for their technical assistance and Joel Dubin and Eric Suess for their assistance with the computer programming. We thank the two reviewers whose insightful comments greatly improved this paper. We thank the California Poultry Industry Federation for its support.

Sources and manufacturers

a. Sigma Chemical Co., St. Louis, MO.
b. BioRad Laboratories, Hercules, CA.
c. New England BioLabs, Beverly, MA.
d. S-plus version 4.0, Mathsoft Inc., Seattle, WA.

References

1. Christensen R: 1990, Log-linear models, 1st ed. Springer-Verlag, New York, NY.
2. Elfadil AA, Vaillancourt JP, Meek AH: 1996, Farm management risk factors associated with cellulitis in broiler chickens in southern Ontario. Avian Dis 40:699–706.
3. Elfadil AA, Vaillancourt JP, Meek AH, Gyles CL: 1996, A prospective study of cellulitis in broiler chickens in southern Ontario. Avian Dis 40:677–689.
4. Elfadil AA, Vaillancourt JP, Meek AH, et al.: 1996, Description of cellulitis lesions and associations between cellulitis and other categories of condemnation. Avian Dis 40:690–698.
5. Gelman A, Carlin J, Stern H, Rubin D: 1997, Bayesian data analysis, 1st ed. Chapman and Hall, New York, NY.
6. Hogg RV, Tanis EA: 1993, Probability and statistical inference, 4th ed. Prentice Hall, Englewood Cliffs, NJ.
7. MacFaddin JF: 1980, Enterobacteriaceae and other intestinal bacteria. *In:* Biochemical tests for identification of medical bacteria, ed. MacFaddin JF, pp. 439–464. Williams and Wilkins, Baltimore, MD.
8. Messier S, Quessy S, Robinson Y, et al.: 1993, Focal dermatitis and cellulitis in broiler chickens: bacteriological and pathological findings. Avian Dis 37:839–844.
9. Murase T, Nakamura A, Matsushima A, Yamai S: 1998, An epidemiological study of *Salmonella enteritidis* by pulsed-field gel electrophoresis (PFGE): several PFGE patterns observed in isolates from a food poisoning outbreak. Microbiol Immunol 40: 873–875.
10. Ngeleka M, Kwaga JKP, White DG, et al.: 1996, *Escherichia coli* cellulitis in broiler chickens—clonal relationships among strains and analysis of virulence-associated factors of isolates from diseased birds. Infect Immun 64:3118–3126.
11. Norton RA, Bilgili SF, McMurtrey BC: 1997, A reproducible model for the induction of avian cellulitis in broiler chickens. Avian Dis 41:422–428.
12. Peighambari SM, Julian RJ, Vaillancourt JP, Gyles CL: 1995, *Escherichia coli* cellulitis—experimental infections in broiler chickens. Avian Dis 39:125–134.
13. Peighambari SM, Vaillancourt JP, Wilson RA, Gyles CL: 1995, Characteristics of *Escherichia coli* isolates from avian cellulitis. Avian Dis 39:116–124.
14. Singer RS, Jeffrey JS, Carpenter TE, et al.: 1999, Spatial heterogeneity of *Escherichia coli* isolated from avian cellulitis lesions in broilers. Avian Dis (in press).
15. Tanner M: 1996, Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions, 3rd ed. Springer-Verlag, New York, NY.
16. Tenover FC, Arbeit RD, Goering RV: 1997, How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists. Infect Control Hosp Epidemiol 18:426–439.
17. Tenover FC, Arbeit RD, Goering RV, et al.: 1995, Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 33:2233–2239.

Appendix

Multinomial versus multiple hypergeometric. Suppose there are i strains on a plate in equal concentration and that a sample of n colonies has been taken from this same plate. Let x_k be the number, out of the n colonies selected, that are of strain type k. A statistical model is required for the vector of counts

$$X \equiv (x_1, x_2, \ldots x_i).$$

If the number of colonies on the plate is large, relative to n, then under the assumptions stated in this manuscript, it is reasonable to assume the vector X has a multinomial distribution with parameter n and with probability vector

$$\left(\frac{1}{i}, \frac{1}{i}, \ldots \frac{1}{i}\right).$$

Alternatively, if the number of colonies of each strain on a plate, N_k, were known, and if N is the total

number of colonies on the plate, it would be reasonable to assume that X is multiple hypergeometric with parameters N (total number of colonies on plate), n (total number of colonies selected), and N_k (total number of colonies of strain k), for $k = (1, 2, \ldots i)$. However, this model is difficult because the values for the N_ks cannot be known, resulting in the assumption of equal numbers for each strain. This would then require a choice for n. It would also require that N is divisible by i, an assumption that may be problematic. Because the use of the multiple hypergeometric distribution is complicated, the multinomial distribution was used for this model, which serves as an approximation to the multiple hypergeometric when N is large relative to n.

Generalization of the multinomial. Let I and J denote random variables that indicate the numbers of actual and observed strains on the plate, respectively. Also let $P(j|i)$ denote the conditional probability that $J = j$ given $I = i$. The values for these probabilities, $P(j|i)$, are calculated as described in the Introduction and are shown in Fig. 1. In the Introduction, the example consisted of two strains growing in equal numbers on an agar plate. In this case, the conditional probability of identifying both strains, assuming three colonies are randomly selected, would be written $P(j = 2|i = 2)$ and was calculated[6] with the binomial distribution. A more general calculation of this probability is made with the multinomial distribution.

Suppose there are i strains on the plate, and a sample of size n colonies is taken. Let $X \equiv (x_1, x_2, \ldots x_i)$ denote the vector of counts where x_k gives the number of colonies out of n that are from each strain k for $k = 1, 2, \ldots, i$. Under the assumptions of the model,

$$X \sim \text{Multinomial}\left(n; \frac{1}{i}, \frac{1}{i}, \ldots \frac{1}{i}\right).$$

$$\text{Let } P_n(x) \equiv \left\{\frac{n!}{\prod_{k=1}^{i} x_k!}\right\}\left(\frac{1}{i}\right)^n; \qquad \sum_{k=1}^{i} x_k = n$$

be the probability function for X. This probability function can then be used to calculate the probability of detecting three strains if there really are three strains and four colonies are selected as follows. In this example, $j = 3$, $i = 3$, and $n = 4$. Then,

$$P(j = 3|i = 3) = P_4(x_1 = 2, x_2 = 1, x_3 = 1) + P_4(x_1 = 1, x_2 = 2, x_3 = 1) + P_4\,(x_1 = 1, x_2 = 1, x_3 = 2),$$

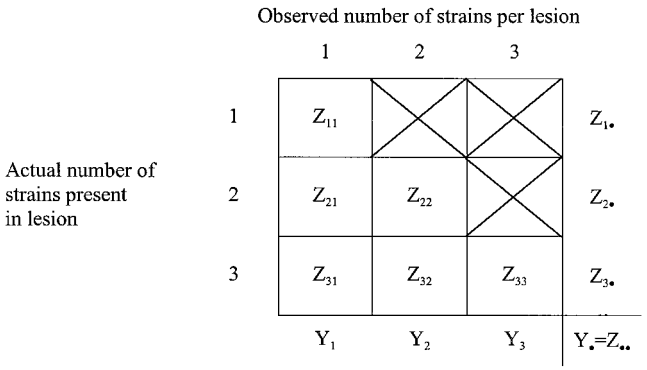
$$P(j = 2|i = 3) = P_4(2, 2, 0) + P_4(2, 0, 2) + P_4(0, 2, 2) + P_4(3, 1, 0) + P_4(3, 0, 1) + P_4(0, 3, 1) + P_4(0, 1, 3) + P_4(1, 3, 0) + P_4(1, 0, 3),$$

and

$$P(j = 1|i = 3) = P_4(4, 0, 0) + P_4(0, 4, 0) + P_4(0, 0, 4).$$

These calculations were used to generate the probabilities of Fig. 1.

Model development. The model was developed as follows. The observed number of strains per lesion was known. However, the number of strains actually present in those lesions was not known. In addition, the true probabilities of a lesion containing one, two, or three strains were unknown. Therefore, there are two sets of unknown parameters. With the use of a Gibbs' sampling routine,[5,15] both of these unknown sets of parameters were estimated simultaneously.



Let Y_1, Y_2, and Y_3 denote the number of lesions in which one, two, or three strains were observed, respectively. Let the subscript ij correspond to the number of strains ("i") actually present and the number of strains ("j") observed. Let Z_{ij} represent the unknown counts (number of lesions) for the true status of the observed data. The counts are unknown because the actual number of strains in the lesion ("i") is unknown. The cells that are crossed out represent impossible outcomes.

Let Θ_i be the prevalence of lesions containing "i" strains; for example, Θ_1 is the prevalence of lesions containing one strain. Denote the vector $(\Theta_1, \Theta_2, \Theta_3)$ as Θ.

Let I and J denote random variables that indicate the numbers of actual and observed strains on the plate, respectively. Denote the conditional probability that $I = i|J = j$ as P_{ij}. With the use of Bayes' theorem, P_{ij} is calculated as:

$$P_{ij} = \frac{\Theta_i P(j|i)}{P(j)}, \qquad P(j) \equiv \sum_{i=1}^{3} \Theta_i P(j|i).$$

The values for $P(j|i)$ are calculated with the multinomial distribution as shown previously. Now, let α_i be the prior guess for the prevalence of lesions containing "i" strains; for example, α_1 is the prior guess for the prevalence of lesions containing one strain. Denote the vector $(\alpha_1, \alpha_2, \alpha_3)$ as α. Let the superscript

attached to parameters in the model denote a particular iterate in the iterative scheme. For example, Θ^0 is the vector at the beginning of the simulation. The prior information α is used as the input values for Θ, namely $\Theta^0 = \alpha$. The prior information for Θ is specified according to the Dirichlet distribution:[5]

$$\Theta^0 \sim \text{Dirichlet}(0.85, 0.1, 0.05);$$

thus the weight parameter associated with the Dirichlet is 1, e.g., the ''prior sample size'' is comparable to a ''data sample size'' of 1, a relatively small weight.

The Gibbs' sampler[5,15] has two stages for problems such as this. Both stages involve sampling from what are called the ''full conditional'' distributions. The process of ''Gibbs sampling'' thus requires successive Monte Carlo draws of observations from the conditional distribution for

$$Z \mid \Theta, Y \qquad \text{(a)}$$

for the ''current'' value of Θ, followed by an observation from the distribution for

$$\Theta \mid Z, Y. \qquad \text{(b)}$$

The sampling for Θ is made with the current value for Z. Again, Y denotes the vector of counts (Y_1, Y_2, Y_3) of the observed data. The conditional distribution of $Z \mid \Theta, Y$ is multinomial and the conditional distribution of $\Theta \mid Z, Y$ is Dirichlet. Details of the sampling procedure from these conditional distributions are shown below. It is straightforward to sample from these distributions with statistical software.[d]

The precise procedure starts with a vector of guesses, Θ^0. With this vector, one samples from the distribution for (a) to obtain Z^1. Then, substituting Z^1 into (b), one samples from the distribution for (b) to obtain Θ^1. This process is then iterated a large number of times.

The probability that there is actually one strain on the plate given that one strain was observed can be imputed by:

$$\begin{aligned} P_{11} &\equiv P(I = 1 \mid J = 1) \\ &= [P(I = 1)P(J = 1 \mid I = 1)] \\ &\quad \div [P(I = 1)P(J = 1 \mid I = 1) + P(I = 2)P(J = 1 \mid I = 2) \\ &\qquad + P(I = 3)P(J = 1 \mid I = 3)]. \end{aligned}$$

At the first iteration, the probability P_{ij} is then:

$$\begin{aligned} P_{ij}^0 &= [\Theta_i^0 \times P(J = j \mid I = i)] \\ &\quad \div \{[\Theta_1^0 \times P(J = j \mid I = 1)] \\ &\qquad + [\Theta_2^0 \times P(J = j \mid I = 2)] \\ &\qquad + [\Theta_3^0 \times P(J = j \mid I = 3)]\}. \end{aligned}$$

With these P_{ij}, the Z_{ij}^1 are obtained by sampling independently from the conditional distributions:

$$(Z_{11}^1, Z_{21}^1, Z_{31}^1) \sim \text{Multinomial}(Y_1; P_{11}^0, P_{21}^0, P_{31}^0)$$

$$(Z_{22}^1) \sim \text{Binomial}(Y_2, P_{22}^0) \text{ and } Z_{32}^1 = Y_2 - Z_{22}^1$$

$$Z_{33}^1 = Y_3.$$

In the second part of the Gibbs sampling routine, the new values Z_{ij}^1 are used to obtain new values for Θ,

$$(\Theta_1^1, \Theta_2^1, \Theta_3^1) \sim \text{Dirichlet}(\alpha_1 + Z_{11}^1, \alpha_2 + Z_{21}^1 + Z_{22}^1, \alpha_3 + Z_{31}^1 + Z_{32}^1 + Z_{33}^1).$$

These new iterates for Θ are then reinserted into the calculations for P_{ij} in order to get new iterates of Z_{ij}. This process was continued for 5,000 iterations. The median of the simulated values was used for a given parameter as a point estimate for that parameter. The 90% Bayesian intervals were placed on the parameters by determining the upper and lower 5% quantiles of the corresponding simulated values for that parameter.[5]

Finally, the probability of correctly identifying all of the strains present in a lesion was calculated for different sample sizes (number of colonies selected). The probability of observing all of the strains present is the sum of three different probabilities and is equal to

$$P(I = J) = \sum_{i=1}^{3} \Theta_i P(J = i \mid I = i).$$

This probability increases as sample size increases. Median estimates and Bayesian intervals were determined as previously described.