Epidemiol. Infect. (2001), **126**, 139–145. Printed in the United Kingdom © 2001 Cambridge University Press

SHORT PAPER

Selection bias in epidemiological studies of infectious disease using *Escherichia coli* and avian cellulitis as an example

R. S. SINGER[1]*, E. R. ATWILL[2], T. E. CARPENTER[1], J. S. JEFFREY[2], W. O. JOHNSON[3] AND D. C. HIRSH[4]

[1] *Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California, Davis, California 95616*
[2] *Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, California 95616*
[3] *Division of Statistics, University of California, Davis, California 95616*
[4] *Department of Pathology, Microbiology and Immunology, School of Veterinary Medicine, University of California, Davis, California 95616*

(*Accepted 12 October 2000*)

SUMMARY

In epidemiological studies of infectious disease, researchers often rely on specific cues of the host, such as clinical signs, as surrogate indicators of pathogen presence. A selection bias would manifest if the specific visual cues used in sampling for the pathogen were not representative of the full range of signs caused by the strains of that pathogen. In our molecular epidemiological studies of *Escherichia coli* associated with avian cellulitis in broilers, we collect carcasses at the processing plant based on visual cues of lesion morphology. Therefore, the objectives of this study were to: (1) explore the potential impacts of selection bias in an application of infectious disease epidemiology, and (2) utilize a validation protocol to assess the potential for selection bias in our molecular epidemiological studies of *E. coli* and avian cellulitis. In two different trials, *E. coli* DNA fingerprints were compared between birds that our observers collected and the birds that the observers missed. Using Fisher's exact tests and simulation models, we determined that the isolates collected by the observers were not significantly different from the isolates missed by the observers ($P > 0{\cdot}60$ in both trials). Our method of selecting birds suspected of having cellulitis did not significantly bias our inferences about the population of *E. coli* associated with cellulitis in the flock. We encourage more investigators to critically assess the relationship of the sample to the target population in epidemiological studies of infectious disease.

A bias in epidemiological studies typically refers to a distortion in the measure of effect between exposures and the outcome due to flaws in either the research design or analysis [1]. In observational epidemiological studies, this distortion can be a result of selection bias, confounding, information bias, which includes recall bias and misclassification, and analytical bias [2, 3]. In this paper, we will focus on selection bias, which generally refers to a spurious estimate of the relationship between exposures and outcome due to the manner in which subjects are selected for study from the target population [1].

In epidemiological studies of infectious disease,

* Author for correspondence: Department of Veterinary Pathobiology, University of Illinois, 2001 S. Lincoln Ave., Urbana, Illinois, USA, 61802.

https://doi.org/10.1017/S0950268801005039 Published online by Cambridge University Press

inferences are often related to specific pathogens and not to the host factors that influence whether an individual is at risk for infection. For example, researchers are often interested in determining the predominant strains of a pathogen involved with a specific condition and how the causal strains are distributed across temporal and spatial scales. Cluster analyses can be used to express the relatedness of isolates sampled in the study. In order to make unbiased inferences about the causal strains, a representative sample of the different strains associated with the condition should be collected. This is often difficult to accomplish because it typically requires a random sample of all hosts in the population that are infected with the pathogen.

Researchers often rely on specific cues of the host as surrogate indicators of pathogen presence. These cues can include specific clinical signs, morphology of gross lesions, biochemical alterations, or radiographic changes. It is possible for different strains of the pathogen being studied to induce variable types and degrees of these clinical manifestations. A selection bias would manifest if the specific visual cues used in sampling for the pathogen were not representative of the full range of signs caused by the strains of that pathogen. Only those strains that produced the specific visual cues would be sampled, and all inferences would then be based on a biased subset of the target population.

We are currently studying avian cellulitis in broiler chickens. This condition is characterized by a diffuse inflammatory reaction in the subcutaneous tissue that results in the complete or partial condemnation of the carcass at processing [4, 5]. Current US losses due to avian cellulitis have been estimated at greater than $80 million per year [6]. Numerous investigators have causally linked the presence of *Escherichia coli* with cellulitis [7–10]. The lesions induced by these *E. coli* can vary considerably in their morphological appearance [5, 11], from mild fibrinous exudate to well-defined, fibrinocaseous plaques. The size of the lesion is extremely variable as is the colour, ranging from a pale cream to golden yellow to brown-yellow with haemorrhage. There can also be considerable variation in the degree to which the overlying skin is affected [5, 11], and some birds have little or no overlying skin involvement.

Currently we are employing pulsed-field gel electrophoresis (PFGE) to assess various aspects of the molecular epidemiology of these cellulitis-associated *E. coli*. For example, we studied the genetic variability of cellulitis-associated *E. coli* within a single flock and the degree of genetic heterogeneity in these *E. coli* across various spatial scales [12]. We also assessed the persistence of these cellulitis-associated *E. coli* over successive flocks that use the same house. In order to sample cellulitis-affected carcasses for *E. coli*, we collect carcasses that we suspect will have cellulitis. Carcasses are collected prior to evisceration on the processing line in order to ensure that the lesions remain uncontaminated [11, 13, 14]. Because we can not directly visualize the lesion in an intact carcass, we use specific changes in the integument of the broiler as indicators of cellulitis. Thus, the cues for selection are visual and may vary from observer to observer.

The strains of *E. coli* that can cause cellulitis are diverse, and thus, there may be variability in the morphology of the lesions that these strains produce. If different strains have different morphologic appearance in the lesions they produce, then some strains may not be considered in a study that selects carcasses based on visual cues. The process of selecting carcasses suspected of having cellulitis based on visual cues may lead to a selection bias and a misrepresentation of the overall target population of *E. coli* that can cause cellulitis in a given flock. Therefore, the objectives of this study were to: (1) explore the potential impacts of selection bias in an application of infectious disease epidemiology, and (2) utilize a validation protocol to assess the potential for selection bias in our molecular epidemiological studies of *E. coli* and avian cellulitis.

Two trials were conducted for this study, each on a single flock from a different ranch. For each trial, two observers collected birds that they suspected would have cellulitis lesions. The observers collected birds from the processing line for a fixed period of time. The birds were collected prior to evisceration so that the lesions would be intact and uncontaminated. During the time period in which the observers were sampling, all birds that the observers missed but that were subsequently condemned for cellulitis by the USDA inspectors at meat inspection were also collected. Therefore, we removed all of the birds in the flock that were identified as having cellulitis during the sampling time period. This sample comprised the cellulitis-affected carcasses that the observers sampled and the cellulitis-affected carcasses that the observers missed.

Each cellulitis lesion was cultured by direct streaking onto MacConkey and blood agar plates. The agar plates were incubated for 24 h at 37 °C. A single, isolated lactose-positive colony was then

https://doi.org/10.1017/S0950268801005039 Published online by Cambridge University Press

randomly selected from the MacConkey plate [15] and identified as *E. coli* by standard biochemical techniques [16].

All isolates were DNA fingerprinted using PFGE. Extraction of genomic DNA and PFGE were performed as per manufacturer's instructions (BioRad Laboratories, Hercules, CA). For extraction of genomic DNA, we utilized the CHEF Bacterial Genomic DNA Plug Kit (BioRad Laboratories, Hercules, CA) as per manufacturer's instructions. Briefly, 62·5 μl of an overnight brain hearth infusion broth culture were centrifuged and washed with TE buffer (10 mM Tris, 1 mM EDTA, pH 8·0). The cells were resuspended in a cell suspension buffer (BioRad Laboratories, Hercules, CA), and an equal volume of molten 2% low-melting-point agarose was added. The combination was mixed and pipetted into the plug mould. Solidified plugs were then treated with a lysozyme solution (BioRad Laboratories, Hercules, CA) for 1 h at 37 °C followed by digestion in a proteinase K solution (BioRad Laboratories, Hercules, CA) at 50 °C overnight. The plugs were then washed four times in 1 × wash buffer (BioRad Laboratories, Hercules, CA), the second of which contained 1 mM phenylmethylsulphonyl fluoride (PMSF). Plugs were then washed in 0·1 × wash buffer for 1 h followed by 1 × restriction enzyme buffer for 1 h. The plugs were digested with 20 U of restriction endonucleases *Not*I and *Sfi*I in separate digestions (New England BioLabs, Beverly, MA) at 37 °C overnight.

PFGE was performed with a 1·2% agarose gel on a CHEF III apparatus (BioRad Laboratories, Hercules, CA) in 0·5 × Trisborate–EDTA buffer (45 mM Tris, 45 mM boric acid, 1 mM EDTA, pH 8·3) at 14 °C and 200 V. A linearly ramped switching time of 5–50 s was applied over 22 h for both *Not*I and *Sfi*I. Three lanes with molecular size markers were included on every gel. The DNA size standards used were bacteriophage lambda ladder (BioRad Laboratories, Hercules, CA) consisting of concatemers with an increment of 48·5 kb.

After PFGE, the gel was stained with ethidium bromide (0·2 μg/ml) and digitized into a computer using the Gel Doc 1000 (BioRad Laboratories, Hercules, CA) and the software program Molecular Analyst v. 1.4.1 (BioRad Laboratories, Hercules, CA). The computerized image was analysed with the Molecular Analyst Fingerprinting Plus software v. 1.12 (BioRad Laboratories, Hercules, CA). Only PFGE fragments larger than 50 kb were evaluated in order to eliminate the potential influence of large plasmids.

Table 1. *Tabulation of the different fingerprint patterns observed in Trial 1 (A) and Trial 2 (B)*

(A) Trial 1

Fingerprint	SAMP1	MISS1
A	7	5
B	3	1
C	1	2
D	1	0
E	0	1
F	0	1
G	0	1
H	0	1
Total	12	12

(B) Trial 2

Fingerprint	SAMP2	MISS2
I	5	7
J	3	4
K	2	2
L	1	0
M	1	0
N	0	1
O	0	1
Total	12	15

SAMP represents the number of isolates of each fingerprint pattern collected by the observers and MISS represents the number of isolates of each fingerprint pattern missed by the observers. Fingerprints A, B, and C were the dominant patterns in Trial 1, and fingerprints I, J, K were the dominant patterns in Trial 2.

In order to assess for the possibility of selection bias, we employed two different analyses. First, we compared the distribution of DNA fingerprint patterns between the sample obtained by the observers (SAMP) and the sample missed by the observers (MISS). We classified each isolate into a DNA fingerprint group using guidelines for interpreting PFGE fingerprint patterns [17, 18]. If isolates had fewer than four band differences, they were considered to be of the same fingerprint type for the purpose of this analysis. The number of isolates in each fingerprint type were then apportioned into a contingency table, comparing the distribution of SAMP to the distribution of MISS. A separate contingency table was designed for each of the two trials (Table 1). The contingency tables were then analysed using a generalization of the Fisher's exact test to compare the vectors of probabilities for the SAMP and MISS

https://doi.org/10.1017/S0950268801005039 Published online by Cambridge University Press

distributions [19]. We performed two different Fisher's exact tests per contingency table. The first test only assessed those fingerprint patterns that had at least two total isolates of each fingerprint pattern (fingerprints A, B and C in Trial 1 and I, J and K in Trial 2). These fingerprint patterns will herein be referred to as the dominant patterns. These comparisons were thus 3×2 analyses. The second test assessed all of the fingerprint patterns that were identified in the trial, and thus, were 8×2 and 7×2 analyses for Trials 1 and 2, respectively.

Because of the relatively small sample sizes in each fingerprint pattern and in order to be more conservative in our assessment of selection bias, the distribution of DNA fingerprint patterns was considered to be significantly different between SAMP and MISS if $P < 0{\cdot}10$. A significant difference in the distribution of DNA fingerprint patterns between SAMP and MISS would imply that the sampled isolates were not representative of the overall population of fingerprints, and consequently, that a selection bias existed in our sampling of *E. coli* from cellulitis lesions in broilers.

In the second type of analysis, we employed a simulation model that was based on a random sampling of the *E. coli* target population in each trial. The purpose of the model was to determine the diversity and frequency of different SAMP distributions that would be obtained under a random sampling scheme from the target population. It would then be possible to assess the likelihood of observing the SAMP DNA fingerprint distribution that we actually obtained. A separate model was created for each trial.

In order to create the model, the SAMP and MISS groups in each contingency table were collapsed into a vector of counts. This then gave the total number of isolates of each fingerprint type that were present in each trial and served as the target population that we were attempting to study. For the purpose of this model, we included only the dominant DNA fingerprint patterns from each trial, and thus, the total number of isolates in the target population for each trial was fixed (n_1 and n_2). In each trial, we had sampled a fixed number of carcasses (SAMP1 and SAMP2), and this number was used as the sample size (x_1 and x_2) for each iteration of the respective model. In each iteration of the model, isolates were sampled at random and without replacement from the target population of dominant fingerprint patterns (n) until the sample size (x) was reached. For example, assume the target population consisted of three dominant fingerprints, A, B and C. Each iteration would then select x isolates from the target population of size n, and this sample would consist of a fingerprints of A, b of B, and c fingerprints of C. We could then determine the proportion of iterations for which $A = a$, $B = b$, and $C = c$. This model was written in Microsoft Access using Visual BASIC® (Microsoft Corp., Redmond, WA), and each trial simulation was run over 10000 iterations.

In Trial 1, 12 birds were selected by the observers (SAMP1 = 12) and 12 carcasses condemned by the USDA inspectors for cellulitis were missed by the observers (MISS1 = 12). In Trial 2, 12 birds were selected by the observers (SAMP2 = 12) and 15 birds were missed (MISS2 = 15). *E. coli* was isolated from the cellulitis lesions in all carcasses. One *E. coli* colony was selected from each lesion and was DNA fingerprinted by PFGE (Fig. 1).

The tabulation of the fingerprint data for each trial is shown in Table 1. In Trial 1, there were three dominant fingerprint patterns (Table 1 A). In Trial 2, there were also three dominant fingerprint patterns (Table 1 B). Again, we refer to a dominant fingerprint pattern as one that was observed in at least two of the isolates in the trial, thus allowing for that fingerprint pattern to be represented in both the SAMP and MISS distributions. Two comparisons were made in each trial between the distribution of fingerprint patterns in the SAMP and MISS isolates. In the first comparison, we considered only the dominant fingerprint patterns, as it is impossible for a fingerprint present in a single isolate to be in both the SAMP and MISS groups. In Trial 1, there was a non-significant difference between the dominant fingerprint distributions in SAMP1 and MISS1 ($P = 0{\cdot}67$). When all isolates were considered, there was still a non-significant difference between the fingerprint distributions in SAMP1 and MISS1 ($P = 0{\cdot}60$). In Trial 2, there was a non-significant difference between the dominant fingerprint distributions in SAMP2 and MISS2 ($P = 1{\cdot}00$). When all isolates were considered, there was still a non-significant difference between the fingerprint distributions in SAMP2 and MISS2 ($P = 0{\cdot}94$).

In the simulation of Trial 1, 11 isolates (representing the 11 SAMP1 dominant isolates) were randomly selected from the target population of 19 dominant isolates of Trial 1 (SAMP1 + MISS1). For the simulation of Trial 2, 10 isolates (representing the 10 SAMP2 dominant isolates) were randomly selected

https://doi.org/10.1017/S0950268801005039 Published online by Cambridge University Press

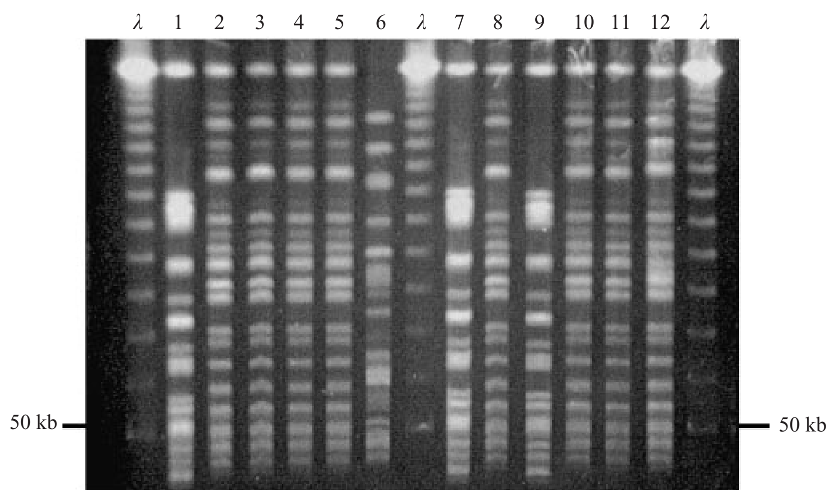

Fig. 1. Pulsed-field gel of DNA obtained from cellulitis *E. coli* in Trial 1 and digested with *Sfi*I. Lanes with λ represent bacteriophage lambda DNA molecular size markers. Only bands larger than 50 kb were considered in the analysis. Isolates 1, 7 and 9 have fingerprint pattern B, isolates 2, 3, 4, 5, 8, 10, 11 and 12 have fingerprint pattern A, and isolate 6 has fingerprint pattern C.

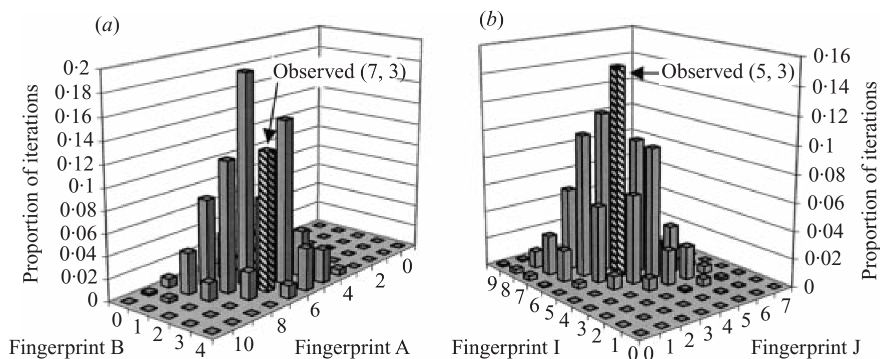

Fig. 2. Results of the simulation models for Trial 1 (*a*) and Trial 2 (*b*). There were three dominant fingerprints in each trial. The counts for the third fingerprint of each simulation are not shown because this count is dependent on the counts of the other two fingerprints. Each bar represents a unique outcome of the model, and the height represents the proportion of the 10 000 iterations that resulted in that specific outcome. The scenario that was actually observed is hatched and labeled.

from the target population of 23 dominant isolates of Trial 2 (SAMP2+MISS2). The different outcomes of each simulation are shown in Figure 2. The figure only shows the counts for two of the three dominant fingerprints in each simulation. This was done because the number of isolates selected from the third fingerprint pattern is dependent on the counts of the other two fingerprint patterns. In Trial 1, the distribution of fingerprint patterns that our observers collected was the third most common simulated outcome. In Trial 2, the distribution of fingerprint patterns that our observers collected was the most common simulated outcome.

Inferences concerning the distribution, diversity and ecology of pathogens in epidemiological studies of infectious disease can be dramatically influenced by the manner in which organisms are sampled. Consequently, infectious disease studies can be greatly affected by selection bias related to the hosts harboring the pathogens under study. We have designed a validation study in order to quantify the amount of selection bias inherent in our molecular epidemiological study of *E. coli* associated with avian cellulitis in broiler chickens.

Approximately equal numbers of carcasses with cellulitis were sampled as were missed by the observers in both trials. The majority of carcasses that were missed had lesions of similar size and morphology to the carcasses that were sampled. The major reason that carcasses were missed was due to the absence of

https://doi.org/10.1017/S0950268801005039 Published online by Cambridge University Press

visible discolouration to the overlying skin. This was the reason that we were concerned about the possibility of a selection bias. If different strains of *E. coli* were to produce different amounts of skin involvement, then we could have underestimated or even missed an entire portion of the *E. coli* strains associated with cellulitis in the flock. Because of the non-significant differences between the SAMP and MISS distributions of DNA fingerprints in both trials, we are confident that our sampling technique is providing a selection of carcasses and *E. coli* isolates that are representative of the target population. In addition, the simulation models demonstrated that our observed outcomes were common in a random sampling design. Based on the dominant fingerprint patterns, we have not missed an *E. coli* strain that could be considered an important component of the cellulitis condition in the flock.

One potential shortcoming of this validation experiment is the limited number of isolates included, resulting in lower power of the study. The argument could be made that a significant difference between the SAMP and MISS distributions would be difficult to detect with only 12–15 isolates of *E. coli* fingerprinted from each group. However, the purpose of this experiment was to detect major departures from a representative sampling scheme, and the number of isolates that were DNA fingerprinted would have given us adequate opportunity to detect these departures. For example, in Trial 1, if our observers had detected three more of the lesions with fingerprint A and missed two more of the lesions with fingerprint B, there would have been a significant selection bias ($P = 0{\cdot}04$). In Trial 2, if our observers had collected two additional isolates of J and missed three more of isolate I, there would have been a significant selection bias ($P = 0{\cdot}07$).

In our sampling of cellulitis-affected carcasses, observers were placed on different evisceration lines and were thus sampling carcasses in parallel. The placement of multiple observers on a single evisceration line could potentially increase the recovery of cellulitis-affected carcasses due to the ability the observers to work in series. However, we did not notice a significant inter-observer variability in the *E. coli* DNA fingerprints recovered (data not shown), and thus, the placement of observers in parallel was probably a more efficient sampling technique.

Many epidemiological studies of infectious disease utilize molecular techniques to differentiate pathogenic isolates. This high degree of resolution has enabled researchers to distinguish closely related isolates within populations. As a consequence, sampling strategies must be carefully designed, taking into account both the distribution of infected hosts within the population as well as the distribution of specific pathogenic strains within the hosts of the population. Random sampling of infected hosts within the target population would clearly be the ideal way in which to ensure that representative strains are collected and identified. Due to problems such as inapparent carriers, diversity of clinical signs and the potential for active infection without clinical lesions, random sampling from all infected individuals can be difficult. Often we rely on cues such as morphologic appearance of lesions in order to obtain our samples. This can potentially lead to selection biases of both hosts and of the pathogens being sampled. A critical assessment of the relationship between the sampled pathogens and the target population would thus be warranted.

ACKNOWLEDGEMENT

This work was supported by a USDA Formula Funds grant and by the Center for Food Animal Health, UC Davis. The authors thank Cara Cooke, Joan Lozano, Karen Tonooka and Shavaun Wolfe for their technical assistance and Dr Bill Sischo for invaluable contributions. The authors wish to thank the School of Veterinary Medicine at the University of California, Davis, and the California Poultry Industry Federation for their support.

REFERENCES

1. Kleinbaum DG, Kupper LL, Morganstern H. Epidemiologic research: principles and quantitative methods. New York: Van Nostrand Reinhold, 1982: 529.
2. Sackett DL. Bias in analytic research. J Chronic Dis 1979; **32**: 51–63.
3. Rothman KJ, Greenland S. Modern epidemiology, 2nd ed. Philadelphia: Lippincott-Raven, 1998: 737.
4. Messier S, Quessy S, Robinson Y, Devriese LA, Hommez J, Fairbrother JM. Focal dermatitis and cellulitis in broiler chickens: bacteriological and pathological findings. Avian Dis 1993; **37**: 839–44.
5. Elfadil AA, Vaillancourt JP, Meek AH, Julian RJ, Gyles CL. Description of cellulitis lesions and associations between cellulitis and other categories of condemnation. Avian Dis 1996; **40**: 690–8.
6. Norton RA. Old principles and new research. Broiler Industry 1998; February: 28–33.
7. Peighambari SM, Julian RJ, Vaillancourt JP, Gyles CL. *Escherichia coli* cellulitis – experimental infections in broiler chickens. Avian Dis 1995; **39**: 125–34.

https://doi.org/10.1017/S0950268801005039 Published online by Cambridge University Press

8. Norton RA, Bilgili SF, McMurtrey BC. A reproducible model for the induction of avian cellulitis in broiler chickens. Avian Dis 1997; **41**: 422–8.
9. Gomis SM, Watts T, Riddell C, Potter AA, Allan BJ. Experimental reproduction of *Escherichia coli* cellulitis and septicemia in broiler chickens. Avian Dis 1997; **41**: 234–40.
10. Jeffrey JS, Chin RP, Singer RS. Assessing cellulitis pathogenicity of *Escherichia coli* isolates in broiler chickens assessed by an in vivo inoculation model. Avian Dis 1999; **43**: 491–6.
11. Onderka DK, Hanson JA, McMillan KR, Allan B. *Escherichia coli* associated cellulitis in broilers: correlation with systemic infection and microscopic visceral lesions, and evaluation for skin trimming. Avian Dis 1997; **41**: 935–40.
12. Singer RS, Jeffrey JS, Carpenter TE, Cooke CL, Chin RP, Hirsh DC. Spatial heterogeneity of *Escherichia coli* isolated from avian cellulitis lesions in broilers. Avian Dis 1999; **43**: 756–62.
13. Peighambari SM, Vaillancourt JP, Wilson RA, Gyles CL. Characteristics of *Escherichia coli* isolates from avian cellulitis. Avian Dis 1995; **39**: 116–24.
14. Ngeleka M, Kwaga JKP, White DG, et al. *Escherichia coli* cellulitis in broiler chickens-clonal relationships among strains and analysis of virulence-associated factors of isolates from diseased birds. Infect Immun 1996; **64**: 3118–26.
15. Singer RS, Johnson WO, Jeffrey JS, et al. A statistical model for assessing bacterial colony selection strategies: A case study of *Escherichia coli* and avian cellulitis. J Vet Diagn Invest 1999; **12**: 118–26.
16. MacFaddin JF. Enterobacteriaceae and other intestinal bacteria. In MacFaddin JF, ed. Biochemical tests for identification of medical bacteria. Baltimore: Williams & Wilkins, 1980: 439–64.
17. Tenover FC, Arbeit RD, Goering RV, et al. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 1995; **33**: 2233–9.
18. Tenover FC, Arbeit RD, Goering RV. How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists. Infect Control Hosp Epidemiol 1997; **18**: 426–39.
19. StataCorp. Stata Statistical Software: Release 6.0. College Station, TX: Stata Corporation, 1999.

https://doi.org/10.1017/S0950268801005039 Published online by Cambridge University Press