Modeling Cumulative Evidence for Freedom From Disease With Applications to BSE Surveillance Trials

Dankmar BÖHNING and Matthias GREINER

This investigation deals with the question of when a particular population can be considered to be disease-free. The motivation is the case of BSE where specific birth cohorts may present distinct disease-free subpopulations. The specific objective is to develop a statistical approach suitable for documenting freedom of disease, in particular, freedom from BSE in birth cohorts. The approach is based upon a geometric waiting time distribution for the occurrence of positive surveillance results and formalizes the relationship between design prevalence, cumulative sample size and statistical power. The simple geometric waiting time model is further modified to account for the diagnostic sensitivity and specificity associated with the detection of disease. This is exemplified for BSE using two different models for the diagnostic sensitivity. The model is furthermore modified in such a way that a set of different values for the design prevalence in the surveillance streams can be accommodated (prevalence heterogeneity) and a general expression for the power function is developed. For illustration, numerical results for BSE suggest that currently (data status September 2004) a birth cohort of Danish cattle born after March 1999 is free from BSE with probability (power) of 0.8746 or 0.8509, depending on the choice of a model for the diagnostic sensitivity.

Key Words: Design prevalence heterogeneity; Diagnostic accuracy; Freedom of disease; Geometric waiting time; Power function.

1. INTRODUCTION

This article focuses on the development of statistical methodology for evaluating a specific population for being free of a particular disease which has recently stirred up considerable interest (Martin, Cameron, Greiner, and Jorgensen 2003; Cameron et al. 2003, 2005). As a particular application we have cattle birth cohorts which are free of BSE for

Dankar Böhning is Professor, Applied Statistics, School of Biological Sciences, University of Reading, Harry Pitt Building, Reading, RG6 6FN, UK (E-mail: *d.a.w.bohning@reading.ac.uk*). Matthias Greiner is Research Professor and Head, International EpiLab, Danish Instiute for Food and Veterinary Research, Møkhøj Bygade 19, DK-2860 Søborg, Denmark (E-mail: *mgr@dfvf.dk*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 11, Number 3, Pages 1–16
DOI: 10.1198/108571106X129117

several European countries for a considerable time period [while it is declining in others (Donelly et al. 1999; Morignat et al. 2002)], and it is of some importance to have a statistical tool available that allows us to evaluate the current evidence for freedom of BSE.

1.1 THE IDEA

Consider a stream of surveillance data obtained from individuals of a population. In terms of the application we think of surveillance data representing the independent testing for BSE in a population of cattle which are obtained in temporal order. Let the binary variable Y_t denote the test result for unit t, where $y_t = 1$ will denote that the individual is positive and $y_t = 0$ otherwise. For $t \neq t'$, it is assumed that Y_t and $Y_{t'}$ are independent. For the time being it is assumed that the test is perfect with sensitivity and specificity both 100%. Interest is in the null hypothesis $H_0 : \pi = 0$, where π is the prevalence parameter of interest. Clearly, the null hypothesis is equivalent to $H_0 : Y_t = 0$, for $t = 1, 2, \ldots$ implying that $P(Y_t = 1|H_0) = 0$ for all times $t = 1, 2, \ldots$, in other words, the probability of a Type I error is zero. Suppose now as the alternative hypothesis that there is some positive (potentially small) prevalence $\pi > 0$. Given the series $Y_1, Y_2, \ldots$, what is the waiting time T such that the first unit is tested positive? Clearly, this waiting time T is defined by sequences

$$1, 01, 001, 0001, 00001, \ldots, \tag{1.1}$$

where the 1 denotes the first unit tested positive. In the hypothesis testing framework, no further diagnostic investigation is required after observing the first event of $Y_t = 1$ and the population is considered not free of the disease. Because the probability for testing positive is π and testing is independent, the associated probabilities for the sequences in (1.1) are

$$\pi, (1-\pi)\pi, (1-\pi)^2\pi, (1-\pi)^3\pi, (1-\pi)^4\pi, \ldots, \tag{1.2}$$

implying that the waiting time T has a *geometric distribution*

$$P(T = t|\pi) = (1-\pi)^{(t-1)}\pi, \tag{1.3}$$

for $t = 1, 2, \ldots$, given that $\pi > 0$. The waiting time is discrete and refers to the number of trials up to and including the last trial, which gives a positive outcome. As a consequence of the geometric distribution we have that

$$P(T > 0|\pi > 0) = \sum_{t=1}^{\infty}(1-\pi)^{(t-1)}\pi = 1, \tag{1.4}$$

implying that with probability one, there exists some positive waiting time for the first unit testing positive, conditional $\pi > 0$. Now, on the contrary, if the entire series $Y_1, Y_2, \ldots$ equals $0, 0, \ldots$, in other words, all units have been tested negative, then it is quite plausible to conclude that $\pi = 0$, for example, the cohort is *disease free*. Because it is impossible to wait for all times t to establish freedom of disease, we are looking for some stopping time

$s < \infty$ such that $P(0 < T \leq s|\pi > 0)$ is *close* to 1. More precisely, given any $\beta > 0$ we are interested in the *smallest stopping time* s such that

$$P(0 < T \leq s|\pi > 0) = \sum_{t=1}^{s}(1-\pi)^{(t-1)}\pi \geq 1 - \beta, \tag{1.5}$$

where β will be small. However, if the Type II error (concluding that $\pi = 0$ when in fact $\pi > 0$) has serious consequences, a choice for β of 0.01 or 0.001 may be more appropriate. If the trial tests positive before reaching s, the population is not disease-free. From the viewpoint of statistical inference, the diagnostic testing can be discontinued, because there is no more interest in the hypothesis testing. If the trial reaches the stopping time s without testing positive, it is taken as evidence for freedom of disease and the trial is terminated. It should be pointed out that the power of this procedure is at least $(1 - \beta)$. We will consider the power

$$\varphi(\pi, s) = P(0 < T \leq s|\pi > 0) \tag{1.6}$$

as a function of the prevalence π and the stopping time s and $\varphi(\pi, s)$ will be the major object of interest here.

The article is organized as follows. In Section 2, we will address the question: Which stopping time is required to reach a predetermined power? Alternatively, given a certain waiting time with associated negative test series, what is the achieved power? To answer these questions, assumptions on the prevalence are necessary. In this context it is important to see that the power function is monotone increasing, both as a function of π and s. Consequently, all that is needed is to specify a threshold value (sometimes called the *design prevalence*) such that prevalences below this value are practically of no interest. In Section 3, nonperfect testing is incorporated. Not all diseased units might be detectable as when not all cattle are BSE-detectable, in particular in the young ages. Therefore, an age-group specific sensitivity lower than one is incorporated. In Section 4, since prevalence is not homogeneous within the population, heterogeneity is included in the modeling. In Section 5, these results are applied to data from Denmark.

2. SOME RESULTS FOR THE POWER FUNCTION

According to (1.5) we have that

$$\varphi(\pi, s) = P(0 < T \leq s|\pi > 0) = \sum_{t=1}^{s}(1-\pi)^{(t-1)}\pi, \tag{2.1}$$

which can be simplified to

$$\begin{aligned} \varphi(\pi, s) &= \sum_{t=1}^{s}(1-\pi)^{(t-1)}\pi = 1 - \sum_{t=s}^{\infty}(1-\pi)^{t}\pi \\ &= 1 - (1-\pi)^{s}\sum_{t=1}^{\infty}(1-\pi)^{(t-1)}\pi = 1 - (1-\pi)^{s}. \end{aligned} \tag{2.2}$$

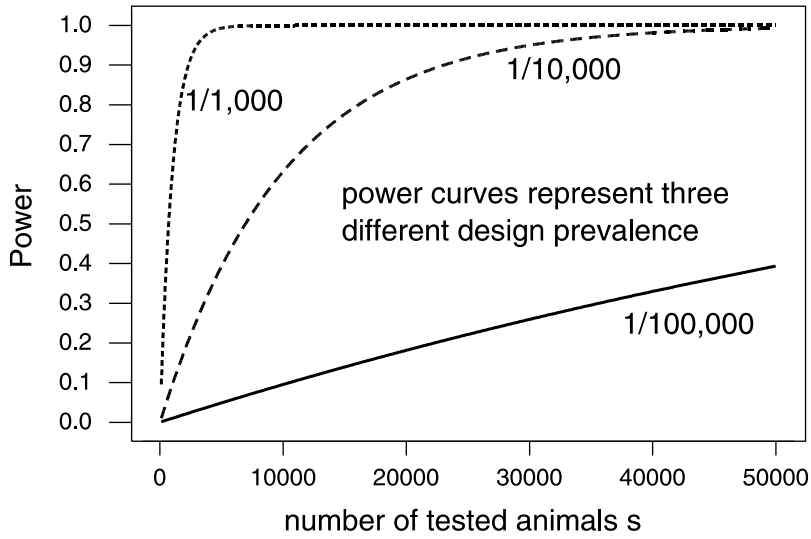

Figure 1. Monotonicity of $\varphi(s|\pi)$ for three different values of the design prevalence, namely 1 in s = 1,000, 10,000, 100,000.

where we used that $\sum_{t=1}^{\infty}(1-\pi)^{(t-1)}\pi = 1$.

Monotonicity of the power function. As direct consequence of the above analytical result $\varphi(\pi, s) = 1-(1-\pi)^s$ we can establish strict monotonicity of the power function.

Result 1. The power function is strictly monotone increasing as a function of s as well as a function of π.

Figure 1 illustrates the monotonicity of $\varphi(s|\pi)$. The power that is reached at stopping time s = 20,000 for design prevalences 1/100,000, 1/10,000, and 1/1,000 is 0.18, 0.86, and 1.0, respectively. In addition, note that if $s_1 < s_2$ and $\pi_1 < \pi_2$, then also $\varphi(\pi_1, s_1) < \varphi(\pi_2, s_1)$ because of the monotonicity in π, and also $\varphi(\pi_2, s_1) < \varphi(\pi_2, s_2)$ because of the monotonicity in s. Consequently, $\varphi(\pi_1, s_1) < \varphi(\pi_2, s_2)$ and $\varphi(\pi, s)$ is strictly monotone as a simultaneous function of s and π.

Waiting time conditional upon design prevalence and desirable power. Because the power function is monotone increasing in the prevalence, any prevalence larger than the chosen prevalence value will reach a power at least as large as the power for the chosen prevalence. This enables the practitioner to overcome the problem of the unknown prevalence parameter. In the following we will think of π as the design prevalence (see, e.g., Wilesmith and Morris 2004). Suppose that the trial should reach a power of at least $1-\beta$. Equating $\varphi(\pi, s)$ to $(1-\beta)$ provides

$$\varphi(\pi, s) = 1-(1-\pi)^s = 1-\beta \tag{2.3}$$

Table 1. Achieved Power $\varphi(\pi, s)$ Given Waiting Time $s = 286, 742$ (in units cattle).

design prevalence: 1 in	*Power*
10,000	1.00000
20,000	1.00000
30,000	0.99993
40,000	0.99923
50,000	0.99677
60,000	0.99160
70,000	0.98337
80,000	0.97224
90,000	0.95866
100,000	0.94316

and the stopping time

$$s = \left[\frac{\log(\beta)}{\log(1-\pi)} \right]$$

having power of at least $(1 - \beta)$. Note that $[x]$ is the smallest integer larger than x. For example, for $\beta = 0.001$ and $\pi = 1/10,000$ we have $s = 6{,}905$. Clearly, the higher the design prevalence, the higher the power. The lower the design prevalence, the longer we have to wait (the more cattle have to be tested) before the trial can be stopped.

Power conditional upon design prevalence and waiting time. On the other hand, we might be interested in determining the power the trial has reached at time s. We can simply calculate $\varphi(\pi, s) = 1 - (1 - \pi)^s$ for various scenarios, as shown in Table 1.

In Table 1 we have computed the power for some scenarios. The waiting time has been chosen according to the number of tested animals in the BSE/TSE database in Denmark. For the application to BSE, the design prevalence follows a recommendation given by the European Commission:

> ". . . since surveillance to date shows BSE prevalence in apparently healthy adult cattle to range from 10 to 100 per million adult bovines in most member States . . ." (EC 2001)

a view adopted here. The prevalence values in the opinion of the European Commission (EC 2001) is reflected in Table 1 (first column).

It is remarkable to see that, even if the design prevalence goes down to 1 in 100,000, the number of units to be tested is sufficient to reach a power of at least 94%, given the current practice of testing for BSE in the European Union.

Design prevalence as a function of the power. In this section we turn the relationship of design prevalence, stopping time s, and power $\varphi(\pi, s) = 1 - (1 - \pi)^s$ around once more. Given a desirable power $(1 - \beta)$ and a size s of the trial, what is the associated design prevalence? That is we have to solve the equation

$$\varphi(\pi|s) = \varphi(\pi, s) = 1 - (1 - \pi)^s = (1 - \beta)$$

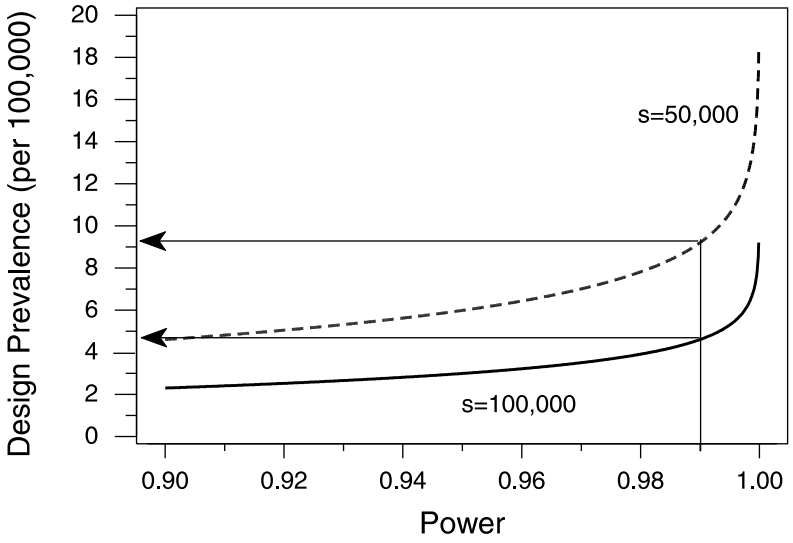

Figure 2. Design prevalence $\pi = \varphi^{-1}(1-\beta|s)$ as a function of the desired power $1-\beta$ for $s = 50{,}000$ and $s = 100{,}000$.

for π given of a particular value of desirable power, conditional upon a specific value of s. The solution is simply

$$\varphi^{-1}((1-\beta)|s) = 1 - \sqrt[s]{\beta}. \tag{2.4}$$

As can be seen in Figure 2, the design prevalence (minimum prevalence that can be detected to reach the desired power) is a monotone increasing function of the power (the higher the desired power, the higher the detectable prevalence, or to put it better: the smaller the desired power, the smaller the detectable prevalence). As can be seen directly from Figure 2, for a desired power of 99% the associated smallest detectable prevalence is 4 in 100,000 for s=100,000 and 9 in 100,000 for $s = 50,000$.

Design prevalence as a function of the stopping time. Finally, suppose that the trial has reached a stopping time s. What is the associated smallest detectable prevalence so that the trial with the given stopping time s achieves a fixed power $1-\beta$? In other words, we are looking for a solution $\pi = \pi(s)$ such that

$$\varphi(\pi, s) = 1 - [1 - \pi(s)]^s = 1 - \beta,$$

where β is a fixed power. The solution is simply

$$\pi(s) = 1 - \sqrt[s]{\beta}$$

identical to (2.4), though the solution is now a function of s. For example, after testing $s = 30{,}000$, a design prevalence of 16 or 24 out of 100,000 can be reached with power of $1-\beta = 0.99$ and 0.999, respectively.

3. EXTENDING THE METHODOLOGY TO NONPERFECT DIAGNOSTIC TESTING

Up to now it is assumed that diagnosis of disease can be achieved without error. It is more realistic to invoke the general diagnostic setting. The probability for the diagnostic test to deliver a positive result will be denoted by π_+ and can be further written as

$$\begin{aligned}\pi_+ &= P(T=1) = P(T=1|D=1)P(D=1) + P(T=1|D=0)P(D=0) \\ &= \alpha\pi + (1-\delta)(1-\pi), \end{aligned} \tag{3.1}$$

where $T = 1$ or $D = 1$ denote that the test is positive or disease is present, respectively. $P(T = 1|D = 1) = \alpha$ is the *test sensitivity* and $P(T = 0|D = 0) = \delta$ is the *test specificity*. For the situation of BSE testing, it can be validly assumed that a BSE-free animal is detected correctly, in other words, we assume that $\delta = 1$ leading to

$$\pi_+ = \alpha\pi \leq \pi. \tag{3.2}$$

We incorporate this modification into the computation of the power function, which leads to

$$P(0 < T \leq s|\alpha, \pi > 0) = \sum_{t=1}^{s}(1-\pi_+)^{(t-1)}\pi_+, \tag{3.3}$$

where $0 < T \leq s$ is again the event that *the waiting time for the first animal testing positive is not above* s. Now, as before

$$\sum_{t=1}^{s}(1-\pi_+)^{(t-1)}\pi_+ = 1-(1-\pi_+)^s = 1-[1-\alpha\pi-(1-\delta)(1-\pi)]^s,$$

where we have used the fact that $\pi_+ = \alpha\pi+(1-\delta)(1-\pi)$ to express the power as function of the prevalence parameter, the sensitivity and the specificity. In the case of BSE, the power computation simplifies to $1-(1-\alpha\pi)^s$.

Capturing the age effect on sensitivity in the case of BSE. Unfortunately, the sensitivity is not identically the same for all cattle as it is dependent on the age as a proxy for the stage of infection. In fact, if the animal is younger than 24 months the test is very unlikely to detect an infection. Therefore, cattle younger than 24 months cannot contribute to the power and are excluded from the analysis. (Alternatively, one might include animals younger than 24 months, but choose a zero-sensitivity for this age-group, leading to the same result as excluding them in the beginning. However, we did not have this option since animals younger than 24 months did not enter into the register). For ages 2 years and above, we have to incorporate the age-dependence of the sensitivity into the modeling. Let $P(T_a > s_a > 0|\alpha_a, \pi > 0) = (1-\alpha_a\pi)^{s_a}$ denote the likelihood for the event that the waiting time T_a for the first animal *from the subpopulation of cattle aged* a *years* testing positive is above s_a, where a goes from age class 1 to A. This means, $(1-\alpha_a\pi)^{s_a}$ denotes the probability

for a Type II error of incorrectly assigning the cohort of cattle as negative based on all infected cattle with age a having a negative test outcome. Given a vector of positive, integer stopping times $\mathbf{s} = (s_1, s_2, \ldots, s_A)'$, we are then interested in the event that *there exists an age-group a such that the waiting time $T_a \leq s_a$*, since in this case the trial would be stopped. Note that inference is made on the basis of all cattle of the birth cohort, regardless of the age at testing. Now, given the prevalence π and a vector of age-specific sensitivities $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_A)'$ we have that the power is provided as

$$\begin{aligned}
\varphi(\pi, \alpha, \mathbf{s}) &= P(\text{there exists at least one age group } a \text{ where the waiting time} \\
&\quad T_a \leq s_a | \alpha, \pi > 0) \\
&= 1 - P(T_a > s_a \text{ for all age groups } a | \alpha, \pi > 0) \\
&= 1 - P\left((T_1, T_2, \ldots, T_A)' > (s_1, s_2, \ldots, s_A)' | \alpha, \pi > 0\right) \\
&= 1 - \prod_{a=1}^{A} P(T_a > s_a | \alpha_a, \pi > 0) = 1 - \prod_{a=1}^{A} (1 - \alpha_a \pi)^{s_a}. \qquad (3.4)
\end{aligned}$$

Note that this is in fact a probability, since $(1 - \alpha_a \pi) \leq 1$ for all ages a and, consequently, $(1 - \alpha_a \pi)^{s_a} \leq 1$ for all ages a, and with this the product $\prod_{a=1}^{A} (1 - \alpha_a \pi)^{s_a} \leq 1$, and it follows that (3.4) is a probability. In addition, the previous result (2.1) is contained as a special case if all age-specific sensitivities coincide.

Before we go ahead to apply these results we consider the effects of incorporating sensitivity into the power computation. Naturally, we expect *a loss of power* due to the loss in ability to detect disease. Indeed, $(1-\alpha_a \pi) \geq (1-\pi)$ for all ages, so that $\prod_{a=1}^{A} (1-\alpha_a \pi)^{s_a} \geq (1 - \pi)^{\sum_a s_a}$ and the following result is achieved.

Result 2. Using the notation of this section and $s = \sum_a s_a$ we have

$$1 - \prod_{a=1}^{A} (1 - \alpha_a \pi)^{s_a} \leq 1 - (1 - \pi)^s. \qquad (3.5)$$

Turning this result around, if sensitivity decreases, more cattle need to be tested to reach the same power.

Using specific sensitivity values in the case of BSE. As mentioned previously, in the situation of BSE, infected cattle can be detected only after considerable time. The sensitivity of the test will depend on the stage of infection. The time from infection to onset of disease is called *incubation time* and can be incorporated into the modeling of age-specific sensitivities. In a key article, Ferguson, Donnelly, Woolhouse, and Anderson (1997) investigated several incubation time models including the three-parameter density given in (3.6)

$$f(a) = \frac{1}{c} \left[\frac{\gamma_2 \exp(-a/\gamma_1)}{\gamma_3} \right]^{\gamma_2^2/\gamma_3} \exp\left[-\frac{\gamma_2 \exp(-a/\gamma_1)}{\gamma_3} \right], \qquad (3.6)$$

where c is the normalizing constant to achieve $\int_0^\infty f(a)da = 1$. Ferguson et al. (1997) provided empirical evidence that the density (3.6) gives a well-fitting distribution and also

Table 2. Age Distribution of Danish Age Cohort and Associated, Potential Sensitivities.[1]

Age group a[2]	*Frequency*	*Sensitivity* α_a	*Sensitivity* α'_a
2	113,197	0.0001	0.0001
3	119,439	0.0114	0.0113
4	50,888	0.1296	0.1196
5	3,218	0.3938	0.3035

[1] Frequency data from the Danish TSE register

[2] Age group *a* refers to years in the interval [*a*, *a* + 1).

derived maximum likelihood estimates using data from the UK for the three parameters as $\gamma_1 = 1.146$, $\gamma_2 = 0.0241$, and $\gamma_3 = 5.71 \times 10^{-4}$, inducing a normalizing constant of $c = 1.134964$. We assume here that, given an infection, the distribution of the time to the point where the disease becomes *detectable* is similar to the distribution of the time to disease onset. One might compute the probabilities for *disease detectability* given infection for discrete ages a as $\int_a^{a+1} f(a')da'$, using the density (3.6). Next, we have to incorporate the distributional character for the time at infection appropriately. Ferguson et al. (1997) presented a model for the time at infection (AI) based on UK data according to which 95% of the BSE cases would have been infected before the age of 1.6. The model has the form

$$\begin{aligned} g(a) &= \gamma_2(\gamma_1 a)^{\gamma_2 - 1} \exp[-(\gamma_1 a)^{\gamma_2}] \left(1 - \exp[-(\gamma_3 a)^{\gamma_2+\gamma_4}]\right) \\ &\quad + (\gamma_2 + \gamma_4)(\gamma_3 a)^{\gamma_2+\gamma_4-1} \exp\left[-(\gamma_3 a)^{\gamma_2+\gamma_4}\right] \left(1 - \exp[-(\gamma_1 a)^{\gamma_2}]\right), \quad (3.7) \end{aligned}$$

with parameter estimates $\gamma_1 = 1.29$, $\gamma_2 = 0.672$, $\gamma_3 = 0.771$, $\gamma_4 = 4.64$. Ferguson et al. (1997) continue to combine the distribution of age-at-infection and incubation time distribution in a basic convolution to yield the distribution of an animal becoming a case. We will denote by $\alpha_a = \sum_{a'=2}^{a} \lambda_{a'}$ the likelihood that an animal becomes a case in the interval from a to $a + 1$ *or before*. It might be argued that this represents a good approximation of the sensitivity and is listed as column three of Table 2 (for the available age groups in the Danish BSE-surveillance data).

Alternatively, it might be argued that, since the animal has lived disease-free up to age a, the likelihood of disease being detectable should be computed *conditional* upon having survived disease-free at age a as

$$\alpha'_a = \frac{\lambda_a}{\sum_{a"=a}^{A} \lambda_{a"}},$$

as listed in numerical values in column four of Table 2. The values of α'_a are lower than those of α_a. It can be shown that this is a general property.

Result 3.

$$\alpha'_a \leq \alpha_a \text{ for all age groups } a = 1, \ldots, A.$$

This result is easily proved using the lemma provided in the Appendix.

4. EXTENDING THE METHODOLOGY TO HETEROGENEITY IN THE DESIGN PREVALENCE

In the previous sections it was assumed that the prevalence is homogeneous in the population of interest. However, it appears to be more realistic to assume that the prevalence varies in the population of interest. In particular, for BSE, prevalence might differ from surveillance stream to surveillance stream.

It is assumed that all covariates are discrete (typically of few categories), so that it is possible to summarize them in covariate combinations or *risk scores* with values r, from 1 to R. Typically, in the case of BSE, we have in mind as risk scores the groups of healthy slaughtered, emergency slaughtered, fallen stock, and clinical suspects. For each of the subpopulations r, there is an associated prevalence π_r. For given stratum specific sensitivities α_a and specificities δ_a, let

$$P(T_{ar} > s_{ar} > 0|\alpha_a, \delta_a, \pi_r > 0) = \{\delta_a + [(1-\delta_a) - \alpha_a]\pi_r\}^{s_{ar}}$$

denote the likelihood for the event that the waiting time T_{ar} for the first unit from the subpopulation of units in *stratum* a and *risk score* r testing positive is above s_{ar}, where a goes from stratum 1 to A and r from risk score 1 to R. In the case of BSE, the strata will be the different age classes of the cattle. Given a matrix of positive, integer stopping times

$$\mathbf{s} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1R} \\ s_{21} & s_{22} & \dots & s_{2R} \\ \dots & \dots & \dots & \\ s_{A1} & s_{A2} & \dots & s_{AR} \end{pmatrix},$$

and a similarly defined matrix of waiting times $\mathbf{T}$, then we are interested in the event that there exists a stratum a and a risk score r such that the waiting time $T_{ar} \leq s_{ar}$, since in this case the trial would be stopped. Now, given a vector of risk score specific prevalences $\pi = (\pi_1, ..., \pi_R)'$ and vectors of stratum-specific sensitivities $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_A)'$ and specificities $\boldsymbol{\delta} = (\delta_1, \dots, \delta_A)'$ we have that the power is provided as

$$\begin{aligned} \varphi(\pi, \alpha, \delta, \mathbf{s}) &= P(\text{ there exists at least one stratum} \\ &\quad a \text{ and a risk score } r \text{ where } T_{ar} \leq s_{ar}|\alpha, \delta, \pi > 0) \\ &= 1 - P(T_{ar} > s_{ar} \text{ for all strata } a \text{ and all risk} \\ &\quad \text{score groups } r|\alpha, \delta, \pi > 0) \\ &= 1 - P\left(\mathbf{T} > \mathbf{s}|\alpha, \delta, \pi > 0\right) \\ &= 1 - \prod_{r=1}^{R}\prod_{a=1}^{A} P(T_{ar} > s_{ar}|\alpha_a, \delta, \pi_r > 0) \\ &= 1 - \prod_{r=1}^{R}\prod_{a=1}^{A} \{\delta_a + [(1-\delta_a) - \alpha_a]\pi_r\}^{s_{ar}}. \end{aligned} \tag{4.1}$$

Table 3. Frequency of Cattle in Healthy Slaughtered and Risk Surveillance Streams by Age Group for the Danish Birth Cohort

Age group[1]	*2*	*3*	*4*	*5*	*All*
Healthy slaughtered	90,511	107,692	46,161	3,029	247,393
Risk group	22,686	11,747	4,727	189	39,349
All	113,197	119,439	50,888	3,218	286,742

[1] Age group a refers to years in the interval $[a, a+1)$.

5. A CASE STUDY: THE DANISH BSE-SURVEILLANCE DATA

In the following we will exemplify the developed model for the surveillance data on BSE in Denmark and will use data coming from the Danish TSE-database. The database is a public register for BSE-testing, controlled by the Danish Veterinary and Food Administration. It is pointed out that currently in Denmark the complete cattle population is included in this registry. There is no random sampling or any other sample selection procedure in place.

The TSE database is a part of GLR/CHR register and it contains information of all animals (including small ruminants) tested for BSE since January 1, 2001. The main purpose of the TSE database is to provide the information required by the EU concerning the number of animals tested and the number of BSE and TSE cases detected. The database also serves as a control system to check whether cattle reported dead in the CHR-register have had a BSE test performed.

For the purpose of this study, the following variables have been used: `Animal-ID`: The official Danish cattle identification number; `Age`: Age of the animal at the time of death in months; `Birthdate`: Date of birth of the animal; `Deathdate`: Date of death of the animal. Identical with the date of sampling; `Submission cause`: provides the reason why the sample was submitted and has the levels: $1 =$ clinical suspect, $2 =$ emergency slaughter, $3 =$ normal slaughter, $4 =$ dead and sent for render, $5 =$ AM-cattle (suspect cases found in the Ante Mortem control by the vets at the abattoir), $6 =$ animal from positive herd, $7 =$ animal from herd under public supervision, $8 =$ tested in connection with export; `Result`: $1 =$ positive, $2 =$ negative. The variable `Result` was used to identify positive cases in the database. There were 13 cases found and the last (youngest) case was born in March 1999.

It is evident from the database that the majority of birth dates of the Danish BSE cases lie between 1996 and 1997, though isolated cases were born in 1998 up to early 1999. Therefore, it appears best to consider as population of interest the *birth cohort* all cattle in the data base with birth date *after March 1999*.

Consider the power according to (2.2) that has been achieved with the Danish surveillance data, where $s = 286{,}742$ is the total of tested aninals in the register. We can simply calculate $\varphi(\pi, s) = 1 - (1-\pi)^s$ for various scenarios, as shown in Table 1.

The power using the specific age distributions of the Danish birth cohort. Due to the young cohort and the considerably reduced sensitivity in young ages, the previously reported

Table 4. Achieved Power $\varphi(\pi, \alpha, \mathbf{s})$ for the Danish Cohort Adjusted for Sensitivity[1] and for Heterogeneity of the Design Prevalence.

	Adjusted for heterogeneity		*Not adjusted for heterogeneity*	
Design prevalence[2] *1 in*	*sensitivity* α_a	*sensitivity* α'_a	*sensitivity* α_a	*sensitivity* α'_a
10,000	0.8746	0.8509	0.6029	0.5693
20,000	0.6459	0.6139	0.3698	0.3437
30,000	0.4994	0.4698	0.2650	0.2448
40,000	0.4049	0.3786	0.2062	0.1899
50,000	0.3398	0.3166	0.1687	0.1550
60,000	0.2925	0.2718	0.1427	0.1310
70,000	0.2566	0.2381	0.1236	0.1134
80,000	0.2286	0.2117	0.1090	0.0999
90,000	0.2060	0.1906	0.0975	0.0894
100,000	0.1875	0.1733	0.0882	0.0808

[1] Age-specific sensitivity estimates α_a, α'_a from Table 2.
[2] The specified design prevalence applies to all surveillance streams when heterogeneity is ignored and to the healthy slaughtered when heterogeneity is accounted for.

rather high power experiences some loss. This will be demonstrated as follows. Let us first consider the age distribution of the Danish cohort as presented in Table 2. Note that the age group of five years is rather sparse, due to the recent nature of the birth cohort. There are no cattle yet above six years. However, the analysis will gain power when later born years are included. As Table 2 shows, the Danish cohort currently experiences the problem that more than 80% of all cattle are in the second and third age group where sensitivity is still low. Consequently, the power, computed as $1 - \prod_{a=1}^{A}(1 - \alpha_a\pi)^{s_a}$ according to (3.4), is dropping down, as Table 4 (columns 4 and 5) shows. In particular, if the design prevalence drops below 1 in 10,000 the power becomes rather small.

Incorporating heterogeneity into the modeling. In the situation of BSE, we can validly assume that the specificity is 100%, for example, $\delta = 1$. Then, (4.1) simplifies to

$$\varphi(\pi, \alpha, \delta, \mathbf{s}) = 1 - \prod_{r=1}^{R}\prod_{a=1}^{A}\{1 - \alpha_a\pi_r\}^{s_{ar}}. \tag{5.1}$$

If we consider the data of the Danish birth cohort, one of the most important covariates is `Reason for Submission`. Almost 87% is recorded as slaughtered when healthy, whereas about 13% is made up by the category `dead and sent for render`. We have therefore grouped everything other than slaughtered healthy into the category `risk group`. Table 3 shows the distribution of the surveillance stream by age group.

For each surveillance stream a value for the design prevalence is needed. The risk ratios reported for Denmark are based on a total of 3 and 2 cases for the years 2002 and 2003, respectively, and are therefore associated with a large statistical uncertainty. Using the data from all EU15 countries (EC 2003, 2004), the combined crude risk ratio and the Mantel-Haenszel risk ratio were established as $RR_{\text{crude}} = 29.3$ and $RR_{\text{MH}} = 18.2$ (95% confidence interval 15.3-21.8) for 2002 and $RR_{\text{crude}} = 27.3$ and $RR_{\text{MH}} = 15.3$ (95% confidence interval 13.0-18.0) for 2003 (Stata Version 8.2; StataCorp, 2003). Based on these empirical results, a risk ratio of 15 was chosen. For the power analysis, the differential

design prevalences were specified such that the minimum value applies to the low-risk group of cattle (HS) and the inflated design prevalence (by factor 15) applies to the high-risk group of cattle. We now apply the result (5.1) to the data of the Danish birth cohort as provided in Table 3 and using a ratio 15 of the design prevalences in risk animals versus healthy slaughtered cattle. The results are given in Table 4 (columns 2 and 3). For comparison we have also included the results, where the heterogeneity in the surveillance stream is *not* incorporated into the modeling (columns 4 and 5). Clearly, the power improves substantially which seems very natural and underlines the importance of sampling from the risk components of the surveillance stream.

6. DISCUSSION

Statistical model. The important underlying assumption, leading to the geometric waiting time distribution, is the *independence* of diagnostic testing. This assumption means that if any two units are tested, the result of one test has no implication on the other. This condition seems reasonable for BSE, because unlike in other diagnostic settings no clustering of infection within a herd occurs for BSE. A potential modification of the distributional model to adjust for dependency would be to allow for a random effects distribution which mixes the geometric distribution over the geometric parameter. This could be developed in a very similar way to other areas where the geometric waiting time distribution has been used, as in fertility studies (Ridout and Morgan 1991; Böhning 2000, p. 177)

Another assumption involved in the geometric distribution is that we are sampling from an infinite population (in other words, the probability for a positive test does not change if a number of units are already tested negative). Though the model could be adjusted for a finite population characteristic, the assumption of an infinite population could be justified by arguing that this is a conservative approach (power will be underestimated rather than overestimated) and that the future population of units (cattle) is unlimited. On the other hand, the adjustment might be necessary for countries with (cattle) populations too small to reach the required power for the specified reference birth cohort.

The model could be also adapted for a null-hypothesis larger than 0. For example, $H_0 : 0 \leq \pi \leq \pi_{DP}$. The major difference is that now a Type I error is possible. From the fixed level for the probability of a Type I error, one can determine a number k, say, of positively tested units, still in agreement with H_0. Then, the sampling space would be the first unit testing positive after having k units positive already. This leads to the negative binomial distribution. The power-analysis would be the same, though technically more complex. For determining the power one could use the right end point of H_0: π_{DP}. However, it should be clearly seen that this generalization leads away from the idea of a disease-free cohort. Therefore, this approach has not been deleveloped any further.

Choice of the statistical power. The question *How much power is enough power?* has no definite answer and the choices will always be also a matter of external, nonstatistical judgement. In experimental studies or clinical trials, typical values for the Type II error (β) are 0.1 or 0.2. This may have important economical (the development of the drug is discon-

tinued) or even public health consequences (e.g., if an unwanted side effect is the subject of the study). The issue of selecting appropriate power values in medical research was reviewed by Muller and Benignus (1992). The authors emphasize the importance of analyzing the power as function of characteristics of the study and conclude that, when ethical and opportunity costs do not preclude it, power should be at least 0.84, and preferably greater than 0.90. The analogy with but also differences between the described statistical test for BSE freedom and experimental studies should be noted. In the latter, the investigator wishes to establish a sample size such that the expected treatment effect can be demonstrated with probability of $1 - \beta$. Whereas the sample size in experimental studies is fixed beforehand, the (BSE) surveillance data can be regarded as a continuously accumulating sample and the choice of a sample size becomes very much a choice of a waiting time to reach sufficient power of the study. In turn, a specified value for the power would determine the waiting time for fixed design prevalence (or vice versa). Section 1 presents a framework for statistical hypothesis testing. Important quantities in the hypothesis testing framework are the sample size (in our case s), the design prevalence (π), and the achieved power ($1 - \beta$), where β denotes the Type II statistical error. In the context of BSE surveillance, β denotes the probability for classifying the age cohort as free of BSE, when in fact it is infected with an unknown prevalence with the lower bound given by π. The achievable power is a function of the principal size of the birth cohort as well as a function of the design prevalence. One could ask the question: *Which design prevalence is reasonable in the light of the size of the birth cohort?* Right now, it appears that in Denmark about 120,000 cattle older than 24 months are tested per birth year. If one assumes as smallest prevalence 1/50,000 we would find, according to Table 4, a power of 17% or 15% (depending which sensitivity model is used). This very low power indicates that surveillance needs to be continued for a considerable amount of time. The OIE terrestrial code specifies a value of 95% for the probability (confidence), to detect foot-and-mouth disease (FMD) or FMD virus infection if present at a defined level of design prevalence (OIE 2004a). This probability has the same interpretation as *power* in the statistical sense. The same level of power is specified in the OIE Code for the demonstration of the absence of infection with highly pathogenic avian influenza (HPAI) virus (OIE 2004b,c). Therefore, a level of 95% seems appropriate for the application to BSE surveillance.

Finally, a consensus value for the required power for BSE surveillance must be found by national and international bodies concerned with risk management. It should be noted that the removal of risk materials from the human and animal food chains is the primary risk mitigation measure, whereas the testing for BSE is one of the corner stones of the geographical risk assessment.

APPENDIX

Lemma 1: Let λ_i be positive real numbers for $i = 1, \ldots, I$ such that $\lambda_1 + \lambda_2 + \cdots + \lambda_I = 1$. Then

$$(\lambda_1 + \cdots + \lambda_i)(\lambda_i + \cdots + \lambda_I) \geq \lambda_i$$

for all i with $1 \leq i \leq I$.

Proof: Let us consider

$$
\begin{aligned}
&(\lambda_1 + \cdots + \lambda_i)(\lambda_i + \cdots + \lambda_I) \\
&\quad = (\lambda_1 + \cdots + \lambda_{i-1})(\lambda_i + \cdots + \lambda_I) + \lambda_i(\lambda_i + \cdots + \lambda_I) \\
&\quad = (\lambda_1 + \cdots + \lambda_{i-1})\lambda_i + \lambda_i(\lambda_i + \cdots + \lambda_I) \\
&\qquad +(\lambda_1 + \cdots + \lambda_{i-1})(\lambda_{i+1} + \cdots + \lambda_I) \\
&\quad = \lambda_i + (\lambda_1 + \cdots + \lambda_{i-1})(\lambda_{i+1} + \cdots + \lambda_I),
\end{aligned}
$$

and the result follows, since $(\lambda_1 + \cdots + \lambda_{i-1})(\lambda_{i+1} + \cdots + \lambda_I) \geq 0$. □

ACKNOWLEDGMENTS

This contribution is a result from the *International EpiLab Project P12* entitled *Development and Evaluation of an Adaptive BSE Surveillance Scheme for Birth Cohorts*. We wish to thank all participants of the project group for advice and support given during the entire study period: Jørgen Nielsen and Mariann Chriél (both Danish Cattle Federation), Anders Stockmarr, Mette M. Andersen, Larry Paisley, Julie Hostrup-Pedersen and Peter Lind (all Danish Institute for Food and Veterinary Research), Preben Willeberg and Helene Rugbjerg (both Danish Food and Veterinary Administration). DB acknowledges the support provided through EpiLab and thanks Anne L. Bisp and René Bødker at the EpiLab for their support in practical matters.

Special thanks go to the Editor and two unknown reviewers for the detailed and constructive comments which greatly helped the authors to provide a timely revision.

[Received TKKK. Revised TKKK.]

REFERENCES

Böhning, D. (2000), *Computer-Assisted Analysis of Mixtures and Applications*, Boca Raton, FL: Chapman&Hall/CRC.

Cameron, A. R., Martin, P. A. J., Greiner, M., and Barfod, K. (2003), "The Use of Scenario-Tree Modelling using Multiple Complex Data Sources to Demonstrate Danish Freedom from Classical Swine Fever," in *Electronic Proceedings of the 10th Meeting of the International Society for Veterinary Epidemiology and Economics (ISVEE)*, Vina del Mar, Chile.

Cameron, A. R., Martin, P. A. J., and Greiner, M. (2005), "Demonstrating Freedom from Disease using Multiple Complex Data Sources: A New Methodology Based on Scenario Trees," *Preventive Veterinary Medicine*, (submitted).

Donnelly, C. A., Santos, R., Ramos, M., Galo, A., and Simas, J. P. (1999), "BSE in Portugal: Anticipating the Decline of an Epidemic," *Journal of Epidemiology Biostatistics*, 4, 277–283.

EC (2001), *Opinion on Requirements for Statistically Authoritative BSE/TSE Surveys.* Adopted by the Scientific Steering Committee (29-30 November 2001). Available online at *http://europa.eu.int/comm/food/fs/sc/ssc/out238_en.pdf*.

——— (2003), *Report on the Monitoring and Testing of Ruminants for the Presence of Transmissible Spongiform Encephalopathy (TSE) in 2002.* Directorate D - Food Safety: Production and Distribution Chain D2 - Biological Risks. Available online at *http://europa.eu.int/comm/food/food/biosafety/bse/annual_report_2002_en.pdf*.

——— (2004), *Report on the Monitoring and Testing of Ruminants for the Presence of Transmissible Spongiform Encephalopathy (TSE) in the EU in 2003, including the Results of the Survey of Prion Protein Genotypes in Sheep Breeds.* Directorate D - Food Safety: production and distribution chain D2 - Biological risks. Available online at *http://europa.eu.int/comm/food/food/biosafety/bse/annual_report_tse2003_en.pdf* .

Ferguson, N. M., Donnelly, C. A., Woolhouse, M. E. J., and Anderson, R. M. (1997), "The Epidemiology of BSE in Cattle Herds in Great Britain. II. Model Construction and Analysis of Transmission Dynamics," *Philosophical Transactions of the Royal Society of London*, 352, 803–838.

Martin, P. A. J., Cameron, A. R., Greiner, M., and Jorgensen, P. H. (2003), "Scenario Tree Modelling of the Danish Diagnostic System to Demonstrate Freedom from Highly Pathogenic Avian Influenza," in *Electronic Proceedings of the 10th Meeting of the International Society for Veterinary Epidemiology and Economics (ISVEE)*, Vina del Mar, Chile.

Morignat, E. C., Ducrot, C., Roy, P., Baron, T., Vinard, J.L., Biacabe, A. G., Madec, J. Y., Bencsik, A., Debeer, S., Eliazsewicz, M., and Calavas, D. (2002), "Targeted Surveillance to Assess the Prevalence of BSE in High-Risk Populations in Western France and the Associated Risk Factors," *Vet Rec*, 151, 73–77.

Muller, K. E., and Benignus, V. A. (1992), "Increasing Scientific Power with Statistical Power," *Neurotoxicology and Teratology*, 14, 211–219.

OIE (2004a), *Guidelines for the Establishment or the Regaining of Recognition for Foot and Mouth Disease Free Country or Zone,* Appendix 3.8.7.of the International Animal HealthCode. Available online at *http://www.oie.int/eng/normes/mcode/en_chapitre_3.8.7.htm*.

——— (2004b), *Highly Pathogenic Avian Influenza*, Chapter 2.7.12 of the International Animal HealthCode. Available online at *http://www.oie.int/eng/normes/mcode/en_chapitre_2.7.12.htm*.

——— (2004c), *Surveillance Systems for Bovine Spongiform Encephalopathy,* Appendix 3.8.4 of the International Animal HealthCode. Available online at *http://www.oie.int/eng/normes/mcode/en_chapitre_3.8.4.htm*.

Paisley, L. G., and Hostrup-Pedersen, J. (2004), "A Quantitative Assessment of the Risk of Transmission of Bovine Spongiform Encephalopathy by Tallow-Based Calf Milk-Replacer," *Preventive Veterinary Medicine*, 63, 135–149.

Ridout, M. S., and Morgan, B. J. T. (1991), "Modelling Digit Preference in Fecundability Studies," *Biometrics*, 47, 1423–1433.

StataCorp (2003), *Stata Statistical Software: Release 8.* College Station: StataCorp LP.

Wilesmith, J. W., and Morris, R. S. (2004), *Development of a Method for Evaluation of National Surveillance Data and Optimization of National Surveillance Strategies for Bovine Spongiform Encephalopathy*, European Union TSE Community Reference Laboratory, Veterinary Laboratories Agency, Weybrige, UK.