ELSEVIER

Preventive Veterinary Medicine 52 (2002) 333–343

PREVENTIVE VETERINARY MEDICINE

www.elsevier.com/locate/prevetmed

Analysis of sampling strategies to substantiate freedom from disease in large areas

M. Ziller*, T. Selhorst, J. Teuffert, M. Kramer, H. Schlüter

Federal Research Center for Virus Diseases of Animals, Institute of Epidemiology, Seestr. 55, D-16868 Wusterhausen/Dosse, Germany

Received 22 September 2000; accepted 20 May 2001

Abstract

In this paper, we deal with the strategies of surveys to substantiate freedom from disease for a certain territory. Infection might not be distributed homogeneously. So, a relatively high within-herd prevalence might be observed while the herd-level prevalence is lower. For this situation, we compare various two-stage sample strategies.

The calculation of appropriate sample sizes becomes quite complicated. The theoretical generalization of the hypergeometric distribution by Cameron and Baldock [Prev. Vet. Med. 24 (1998) 1] introduces a simple way to evaluate multi-stage sample sizes while regarding real-test properties. We demonstrate the theoretical foundations of these calculations. These principles open up the possibility of optimizing costs or other relevant variables, by choosing the appropriate sample strategy (each of which ensures the same α-level for the first stage). In addition, we evaluate the statistical power of the complete strategies under consideration.

Furthermore, we apply our theoretical results to a data example of *Brucella melitensis*. We used the herd-size situation in Germany, characterized by many small sheep holdings and only a few large ones. The consequences of real-test properties on sample sizes and on the applicability of several strategies are discussed.

Keywords: Survey; Sample size; Statistical power; Diagnostic test

1. Introduction

Recently, regulations have been introduced to determine whether an area or a certain holding can be declared as being free from a specific disease. Several different situations have to be distinguished from each other. The ‘‘disease-free’’ status can be granted initially as a result of control measures, reinitiated after rehabilitation or continue after a period of

* Corresponding author. Tel.: +49-33979-80182; fax: +49-33979-80200.
E-mail address: mario.ziller@wus.bfav.de (M. Ziller).

0167-5877/02/$ – see front matter
PII: S0167-5877(01)00245-8

observation. In some instances, ''disease-free'' requires that not a single infected animal be present. In still other instances, a certain percentage-based sampling survey is prescribed. Also, clear goals have been formulated (sometimes in the form of prevalence-thresholds). The typical aim of sampling surveys should be the possible detection of a certain prevalence level at a definite time, depending on the disease and the specific situation.

Infection might not be distributed homogeneously. So, when infection occurs, a relatively high within-herd prevalence of infected animals might be observed while the herd-level prevalence remains lower. An example is in Fig. 1: infected holdings can be clustered but only a low herd prevalence exists. However, a much higher prevalence might be expected instead within an infected holding. For this situation, Cameron and Baldock (1998b) recommended two-stage sampling. Such strategies are flexible enough to permit monitoring and observation programs with very different parameters (prevalence levels and error levels).

To prove that a territory was free from disease a certain time ago, it is usually sufficient to show that no consequences of the supposed infection now occur. For contagious diseases, maintaining different low-level prevalence-thresholds (amongst herds as well as within herds) is sufficient to indicate freedom during a certain time. If animals were already infected, the prevalence would soon exceed those thresholds. In general, the threshold for testing within herds should be greater than that for testing at the herd-level.

The calculation of sample sizes of multi-stage sampling strategies is more complicated. The classical standards (e.g. collected in the tables of Cannon and Roe (1982)) were developed for easy, one-stage samples and perfect tests. As an application of the theorem of total probability (Ash, 1999), Cameron and Baldock (1998a) published a formula suitable for sample-size calculations for single-stage samples, considering sensitivity and specificity. Furthermore, this theoretical generalization of the hypergeometric distribution introduces a simple way to evaluate multi-stage sample sizes with real-test properties.

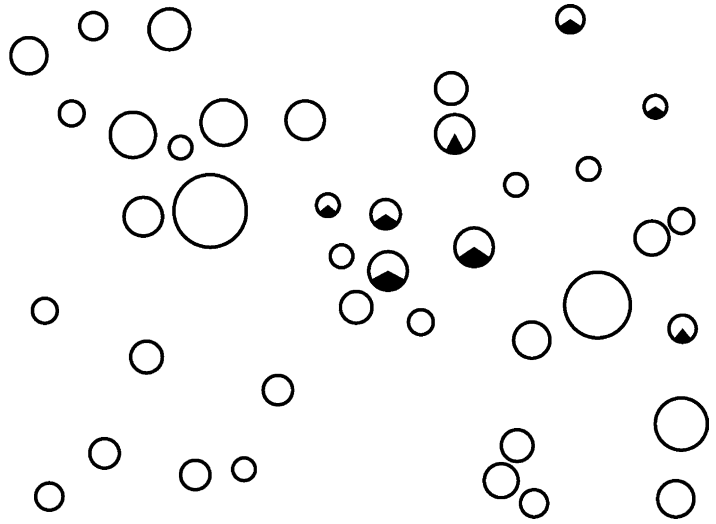

Fig. 1. Scheme of an infected territory. Area of the circles: herd sizes; black segments: infected percentages.

(Generally applying the hypergeometric distribution, however, was problematic until more-powerful computers were available.) In this paper, we demonstrate the theoretical foundations of these calculations. In this, each stage has to be regarded separately. The ''interface'' between both stages is characterized by the test-error probabilities of the following step (Ziller et al., 1999).

We use a data example of *Brucella melitensis.* We take special note of the herd-size situation in Germany, characterized by many small sheep holdings and only a few large ones (a distribution typical for different animals in various countries). Whilst Cameron and Baldock (1998b) investigated the case of homogeneous herd-size distribution, the simulation study of Selhorst et al. (1999) was applied to the more-general situation.

1.1. Two-stage sample strategies

In the first stage, a random sample of holdings is selected. In the second stage, each of these holdings must be determined to be infected or not. What is done in this second stage characterizes the strategy. We compare three strategies for the second stage: cluster samples (Teuffert and Lorenz, 1995), individual samples (Ziller et al., 1999), and limited samples (Cameron and Baldock, 1998b; Selhorst et al., 1999; Ziller et al., 1999) (Fig. 2).

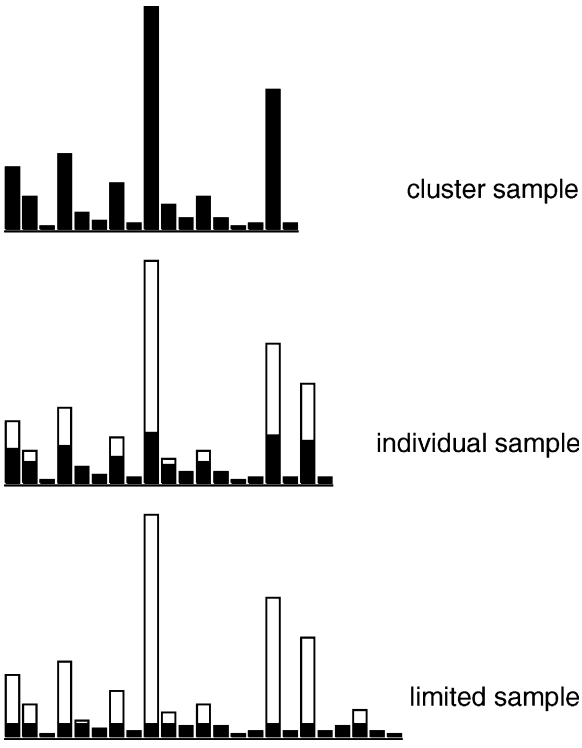

Fig. 2. Scheme of sample strategies. Total bars: herd sizes of sampled herds; black areas: sample sizes within sampled herds.

1.1.1. Cluster sample

Cluster sampling involves all animals within the herd (i.e. within the selected clusters). By assuming a perfect test, we can be sure of deciding the herd's status correctly.

1.1.2. Individual sample

Individual sampling involves samples of the selected herds relative to a pre-defined prevalence-threshold and thus depends on the herd size. If the herd is infected, test-positive animals should be found in an appropriate sample. This sample size has to be calculated based on the size of the herd so as to ensure a chosen error level (e.g. $\alpha_2 = 0.05$) for each tested herd.

1.1.3. Limited sample

Limited sampling involves a pre-fixed number of animals per each selected herd irrespective of the herd size (if a holding has fewer animals, all of them have to be tested). In the case of limited sampling, knowledge of the herd-size distribution is essential so that the necessary number of herds can be calculated to ensure the corresponding α_1-level for the first stage. Depending on the chosen sample size within the holdings and the particular herd-size distribution, the total number of tested animals can differ from situation to situation in this strategy. Therefore, this strategy can be chosen to optimize a given external criterion.

1.2. Real-test properties

All diagnostic tests actually have sensitivities and specificities <1 (methodical errors cannot be excluded completely). Therefore, in practice, it is usual to verify positive test results when performing large surveys. For sample-size calculations, we summarize all information from the whole process of testing in series into one sensitivity and one specificity (and, we assumed independence between tests).

2. Materials and methods

2.1. Sample sizes

In general, the sample size (n) depends on the error probability (α), the prevalence-threshold (π) to be recognized, the population size (N), and the sensitivity (Se) and specificity (Sp). n is the minimum sample size, for which the probability of having no test-positive result within the sample is $\leq\alpha$. M denotes the assumed number of true-positives (corresponding to π) within the population. The summation index (m) represents the possible number of true-positives within the sample.

$$P(T^+ = 0) = \sum_{m=\max(0,n-N+M)}^{\min(n,M)} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} (1 - \mathrm{Se})^m \mathrm{Sp}^{n-m} \leq \alpha. \quad (1)$$

For control programs, a test-positive result has to be treated as if true-positive. To calculate sample sizes, one has to set Sp = 1 in Eq. (1). For these reasons, a sample survey of this kind only makes sense if $\pi \gg 1 - \mathrm{Sp}$. (Otherwise, there is a great chance of having false-positive results within the sample, and of unnecessary control measures being invoked.)

Now, we describe the theoretical basis for exactly calculating the sample sizes for two-stage sampling plans. Let an area be regarded as free from disease when the prevalence of infected herds is $<\pi_1$. We now consider the procedure of checking a holding for infection as being a ''whole'', and call it ''herd-test''. We estimate the sensitivity Se_1 of this herd-test, depending on the specific sample sizes (n_2) within the herds assumed to be infected, and the known herd-size distribution. Let the summation index (i) denotes the possible herd sizes.

$$\mathrm{Se}_1 = 1 - \sum_{i=1}^{\infty} P(T^+ = 0 | n_2 = i) \cdot P(n_2 = i). \tag{2}$$

The estimation of the herd-test sensitivity is the key to evaluating the complete sampling plan.

2.1.1. Cluster-sample A

All of the N_2 animals within the herd are tested. The herd is regarded as infected if ≥ 1 animal is test-positive (this is the usual, more-stringent version of understanding cluster samples). Thus, the herd-test sensitivity at least equals the sensitivity (Se_2) of testing a single animal. Indeed, Se_1 might be even greater than Se_2 if more than one animal is infected—but for sample-size calculations, we have to assume the worst case:

$$n_2 = N_2, \qquad \mathrm{Se}_1 = \mathrm{Se}_2. \tag{3}$$

2.1.2. Cluster-sample B

All of the N_2 animals within the herd are tested. The herd is regarded as infected if the prevalence of infected (true-positive) animals $\geq \pi_2$. Denoting the smallest integer $\geq x$ by $\lceil x \rceil$ yields:

$$n_2 = N_2, \qquad \mathrm{Se}_1 = 1 - \sum_{i=1}^{\infty} (1 - \mathrm{Se}_2)^{\lceil \pi_1 \cdot N_1 \rceil} \cdot P(N_2 = i). \tag{4}$$

2.1.3. Individual sample

The sample size at the level of animals (within herd) has to be determined individually from the parameters of the herd. The herd-test sensitivity is complementary to the error level of the individual within-herd test. We assume a sampling procedure within the holding to ensure a prevalence $<\pi_2$ at an error level α_2. For the case of $\mathrm{Se}_2 \geq 1 - \alpha_2$, this yields:

$$\mathrm{Se}_1 = 1 - \alpha_2. \tag{5}$$

Indeed, Se_1 might be even greater than $1 - \alpha_2$ because the sample sizes within the herds have to be rounded up, but again the worst case is assumed. In the case of $\mathrm{Se}_2 < 1 - \alpha_2$, it is impossible to guarantee an error probability of α_2 (especially for small herds). So the

assumptions of individual samples are violated; the actual α_2 has to be calculated for the specific situation.

2.1.4. Limited sample

This case is more complicated. The probability of a false-negative determination on a holding can be calculated only by using the formula of total probability. The herd-size distribution is needed. Supposing the same prevalence-threshold (π_2) as for the individual sample, let k be the fixed sample size (equal for each tested holding). If a holding has fewer animals, all of them have to be tested.

The error of the first kind $(1 - \mathrm{Se}_1)$ of the herd-test is the sum over all possible population sizes (i) of the conditional probabilities of choosing a sample of n_2 test-negative animals, given a population size i, multiplied by the probability of a holding having a total of i animals (from the herd-size distribution).

$$\mathrm{Se}_1 = 1 - \sum_{i=1}^{k-1} P(T^+ = 0 | N_2 = n_2 = i) \cdot P(N_2 = i) - P(T^+ = 0 | N_2 \geq k \wedge n_2 = k) \cdot P(N_2 \geq k). \quad (6)$$

For each choice of k, we will get a new sampling plan. These will differ from one another by the total number of animals to be tested and therefore by the total costs or other criteria. Thus, an optimal sample strategy can be found by choosing an appropriate limit k.

2.2. Power calculations

Sample-size calculations that only consider the α-error level are not sufficient for the evaluation of sample strategies (Ziller et al., 2000). The statistical power of the sampling procedure (regarded like a statistical test) is also an important aspect. Thus, we must estimate the probability of recognizing a truly disease-free territory as being free. In all power calculations, we have to assume that no animal within the territory is infected. Given the actual specificity (Sp_1) of the herd test and the number (n_1) of herds to be tested, the statistical power of the whole sampling procedure results in:

$$\mathrm{Sp} = {\mathrm{Sp}_1}^{n_1}. \quad (7)$$

The estimated specificity of the herd test depends on the chosen strategy. Given the specificity (Sp_2) of the diagnostic procedure, the power of the sampling investigation depends only on the total number of animals to be investigated. Simply spoken, the statistical power is lower when the total sample size is larger.

$$\text{Cluster sample}: \quad \mathrm{Sp}_1 = \sum_{i=1}^{\infty} {\mathrm{Sp}_2}^{N_2} \cdot P(N_2 = i), \quad (8)$$

$$\text{Individual sample}: \quad \mathrm{Sp}_1 = \sum_{i=1}^{\infty} {\mathrm{Sp}_2}^{n_2} \cdot P(N_2 = i), \quad (9)$$

$$\text{Limited sample}: \quad \mathrm{Sp}_1 = \sum_{i=1}^{k-1} {\mathrm{Sp}_2}^{i} \cdot P(N_2 = i) + {\mathrm{Sp}_2}^{k} \cdot P(N_2 \geq k). \quad (10)$$

Table 1
Calculated parameters for strategies with perfect specificity, and various animal-level sensitivities (Se_2)

Sampling strategy	Sample size (size limit, respectively, within a holding)	Herd-test sensitivity (%)			Number of herds to be tested			Expected number of animals to be tested $\times 10^3$		
		Se_2 1.00	0.95	0.90	Se_2 1.00	0.95	0.90	Se_2 1.00	0.95	0.90
Cluster A	N_2	100	95	90	1476	1553	1640	50	52	55
Cluster B	N_2	100	95	90	1476	1549	1631	50	52	55
Individual	n_2	99	95	90	1488	1553	1635	30	32	35
Limited	15	87	82	78	1706	1796	1895	14	14	15
	14	86	81	77	1725	1816	1917	13	14	15
	13	85	80	76	1747	1839	1941	13	14	14
	12	83	79	75	1772	1865	1969	13	13	14
	11	82	78	74	1802	1897	2003	12	13	14
	10	80	76	72	1839	1935	2043	12	12	13
	9	78	74	71	1884	1983	2093	11	12	13
	8	76	72	69	1937	2039	2152	11	12	12
	7	74	70	66	2004	2110	2227	10	11	12
	6	71	67	64	2093	2203	2325	10	10	11
	5	66	63	60	2222	2339	2469	9	10	10
	4	61	58	55	2434	2562	2705	9	9	10
	3	52	50	47	2831	2981	3146	8	8	9
	2	39	37	35	3772	3971	4191	7	8	8
	1	21	20	19	6893	7256	7659	7	7	8

2.3. Data example

To demonstrate our results, we have chosen an imaginary survey to substantiate freedom from *B. melitensis* for Germany. According to the EC-directive 91/68/EEC (1991), for maintaining the status ''free of infection'' for a Member State it is sufficient to prove at the $\alpha_1 = 0.05$ level that the prevalence of infected herds is <0.002. Infection-free herds means free of clinical signs and free of antibodies.

According to the Federal Office Statistics, 68,767 holdings existed in Germany at the end of 1996 (2,314,870 sheep). About 64% of them contained 10 animals at the most, and only 246 holdings with more than 1000 animals could be found. Therefore, a cluster-sample procedure is not only very expensive—but also inefficient and unfair to the holders of large herds.

2.4. Computing

All three sample strategies under consideration were evaluated using the example *B. melitensis* for Germany. Therefore, the error level $(\alpha_1 = 0.05)$ and the prevalence-threshold $(\pi_1 = 0.002)$ at the herd-level were given. The following parameters for the second stage within the holdings were fixed: prevalence-threshold $\pi_2 = 0.01$, maximum error probability $\alpha_2 = 0.05$. We calculated herd-test sensitivities, and numbers of herds and animals to be tested for all three testing strategies, with animal-level sensitivity $Se_2 = 1.00, 0.95, 0.90$ (each with $Sp_2 = 1.00$), and in addition with animal-level specificity $Sp_2 = 0.9999, 0.99999$ (each with $Se_2 = 0.95$).

Table 2
Calculated power values for strategies with $Se_2 = 0.95$

Sampling strategy	Sample size (size limit, respectively, within a holding)	Power of the whole sampling survey		
		Sp_2 1	0.99999	0.9999
Cluster A	N_2	1.000	0.5936	0.0061
Cluster B	N_2	1.000	0.5944	0.0061
Individual	n_2	1.000	0.7255	0.0409
Limited	15	1.000	0.8656	0.2362
	14	1.000	0.8686	0.2445
	13	1.000	0.8718	0.2536
	12	1.000	0.8752	0.2636
	11	1.000	0.8787	0.2746
	10	1.000	0.8826	0.2868
	9	1.000	0.8867	0.3004
	8	1.000	0.8911	0.3157
	7	1.000	0.8958	0.3329
	6	1.000	0.9010	0.3526
	5	1.000	0.9066	0.3752
	4	1.000	0.9127	0.4011
	3	1.000	0.9191	0.4299
	2	1.000	0.9250	0.4586
	1	1.000	0.9300	0.4840

All the calculations were performed with S-PLUS 2000 (©1999 Mathsoft, Seattle, USA). The corresponding script will be made available on request.

3. Results

If $Se_2 < 1$ then Se_1 becomes smaller, too, and the number of investigated herds must be larger to ensure the same error probability for the whole sampling plan (Table 1).

In fact, no diagnostic procedure can reach a perfect specificity. What would happen when the specificity of the test procedure within the herds actually is below 1 ($Sp_2 < 1$)? Effort and expense remain the same, but the statistical power (ability to recognize freedom from disease) of the whole sampling survey instead becomes less when the chance of false-positive test results increases. Table 2 demonstrates this effect for $Sp_2 = 0.99999$ and 0.9999, respectively (for the example $Se_2 = 0.95$ was used).

The considered specificities seem to be rather high. But even for these values, the statistical power of the whole sampling procedure is seriously affected. In practice, specificity values may be even smaller.

4. Discussion

For sample-size calculations for a multi-stage sampling strategy, all the stages have to be regarded separately. Sensitivity and specificity for the appropriate stages have to be estimated. Only after that sample sizes can be calculated.

4.1. Cluster sample

A cluster-sample strategy results in the statement of highest certainty, but also in highest costs (much larger numbers of tests than with the other strategies). So this strategy can be recommended only when political requirements make it indispensable.

4.2. Individual sample

The individual sample is much more efficient than cluster samples and allows a result at an acceptable confidence level for each investigated holding. This advantage could make the holders of large herds more willing to collaborate in an extensive survey.

The strategy can be recommended when $Se_2 \geq 1 - \alpha_2$ holds true and when statements for each investigated holding are desired. Nevertheless, the logistic effort is higher than that for cluster sampling and it cannot be a primary aim of most large surveys to yield specific information for each holding.

4.3. Limited sample

The strategy of limited samples is the least expensive. Compared to the other strategies, the limited samples strategy requires more holdings to be tested, but the total number of

animals necessary to be tested is much lower. For this reason, the statistical power of the whole survey becomes higher too. On the other hand, although the α_1-level for the first stage is guaranteed, the herd-level sensitivity is lower, and a generally reliable statement for single holdings is impossible here. The strategy of limited samples is strongly recommended for large surveys when statements for each investigated holding are not required.

The limited-sample strategy permits the arbitrary choice of the sample limit of animals to be tested within one holding. Therefore, it can be adjusted to allow optimizing a secondary cost function. This represents another advantage of limited samples.

4.4. Surveys

We considered several strategies of surveys to ensure a given herd-level prevalence-threshold in a certain territory. Animal-level sensitivity and specificity of the diagnostic tests influence the sample size necessary to guarantee the pre-fixed α-level, and the statistical power of the survey as well. Two simple consequences are obvious. A decrease of the sensitivity requires a slightly larger sample size to compensate for possible false-negative results, and therefore causes a little lower statistical power. A decrease of the specificity, instead, does not affect the sample size itself—but it leads to a serious reduction of the statistical power because of the risk of false-positive results.

In our presented example, even a specificity of $Sp_2 = 0.9999$ is much inadequate. In this case, the probability of a false-positive decision after the whole survey is greater than that of accurately detecting a free territory as being free. So enhancing the specificity of the diagnostic procedure is essential for effectively performing large surveys. Though resulting in a relatively low herd-test sensitivity, introducing the strategy of limited samples will reduce the general effort.

A survey of the kind considered is not model based; it is closely related to legal regulations for a territory to maintain the status ''free from disease''. It should discover a potential infection early to prevent a further spread. So frequency and intensity of the investigations must depend on the specific disease and the situation. Certainly, other questions must also be considered. When regulations or political restraints require large surveys, a cost-optimized strategy ensuring all predefined parameters would be the best solution. Last but not the least, ethical questions may be dominant. When animals have to be slaughtered for testing, a strategy should be chosen which minimizes the total sample size.

References

Ash, R.B., 1999. Probability and Measure Theory. Harcourt Academic Press, San Diego, CA.

Cameron, A.R., Baldock, C., 1998a. A new probability formula for surveys to substantiate freedom from disease. Prev. Vet. Med. 24, 1–17.

Cameron, A.R., Baldock, C., 1998b. Two-stage sampling in surveys to substantiate freedom from disease. Prev. Vet. Med. 24, 19–30.

Cannon, R.M., Roe, R.T., 1982. Livestock Disease Surveys. A Field Manual for Veterinarians. Bureau of Rural Sciences, Department of Primary Industry. Australian Government Publishing Service, Canberra.

EEC, 1991. Council directive on animal health conditions governing intra-community trade in ovine and caprine animals, 91/68/EEC.

Selhorst, T., Teuffert, J., Staubach, C., Ziller, M., Schlüter, H., 1999. Entwicklung eines effizienten Stichprobenplanes am Beispiel der Schaf- und Ziegenbrucellose *Brucella Melitensis* (Development of an efficient sample strategy to substantiate freedom from *Brucella melitensis* in sheep). Z. Agrarinformatik 7, 27–32.

Teuffert, J., Lorenz, R.J., 1995. Die Erarbeitung von Stichprobenplänen in Tierbeständen mit dem Ziel des Nachweises der Seuchenfreiheit von Gebieten am Beispiel der Schaf- und Ziegenbrucellose (Elaborating sampling plans for holdings in order to prove freedom from epidemics at the example of *Brucella melitensis*). Report of the Federal Research Centre of Virus Diseases of Animals. Institute of Epidemiology.

Ziller, M., Selhorst, T., Teuffert, J., Schlüter, H., 1999. Sample strategies to substantiate freedom from disease. A theoretical approach. In: Goodall, E.A., Thrusfield, M.V. (Eds.), Proceedings of the SVEPM, Bristol, UK, March 24–26, 1999, pp. 44–52.

Ziller, M., Kramer, M., Schlüter, H., Selhorst, T., Teuffert, J., 2000. Zum Einfluß der Genauigkeit diagnostischer Tests auf die Güte verschiedener Stichprobenstrategien im Zusammenhang mit dem Nachweis der Freiheit von einer Krankheit (On the influence of the accuracy of diagnostic tests on the power of various sample strategies to substantiate freedom from disease). In: Proceedings of the 46th Biometric Colloquium of the German Region of the International Biometric Society, Rostock, Germany, March 20–23, 2000, p. 72.