EUROPEAN COMMISSION
HEALTH & CONSUMER PROTECTION DIRECTORATE-GENERAL

Directorate C - Scientific Opinions

OPINION of the SSC ON

DESIGN OF A FIELD TRIAL FOR THE EVALUATION OF NEW RAPID BSE POST MORTEM TESTS

ADOPTED ON 22 FEBRUARY 2002

Keywords: Bovine Spongiforme Encephalopathy, rapid tests, evaluation, field trial

1. Introduction:

Following a call for expression of interest several parties indicated their interest to participate with their new rapid BSE post mortem tests in an EC evaluation exercise. In a pre-selection, 5 of these tests were selected for participation in a laboratory evaluation exercise conducted by the Institute of Reference Materials and Methods (IRMM).

The results of this laboratory evaluation exercise were discussed by an expert group on 12 December 2001. On 11 January 2002 the Scientific Steering Committee (SSC) recommended that these tests should undergo an evaluation under field conditions (field trial) prior to approval. The Commission Services invited the SSC to prepare a design for such a field trial.

In order to gather as much experience in respect of rapid BSE testing as possible, an expert meeting was held on 11 and 12 February 2002, aiming to join the scientific and technical experience available at the level of the TSE/BSE ad hoc group, all National Reference Laboratories (NRL) and elsewhere (e.g. Swiss Veterinary Administration, biostatisticians). No experts from France, The Netherlands, Greece and Finland could attend.

2. Purpose

According to EU legislation all slaughtered cattle over the age of 30 months have to be tested using one of the approved[1] "rapid BSE tests". In addition, a certain sample size of fallen stock over 24 months of age as well as all emergency slaughtered cattle over 24 months of age have to be subjected to an approved rapid test. Presently, test kits of three manufacturers are in use after having been evaluated and subsequently approved by the EU Commission. In the meantime, five new rapid tests have been developed and have taken part in an EC laboratory evaluation exercise according to a protocol defined by the European Commission. All of the test results showed the potential to fulfil minimum parameter requirements. In considering the evaluation report the SSC, in agreement with the IRMM, concluded that each of the new rapid tests should be subjected to an additional field trial comparing the new rapid test with already approved tests. The performance of any new rapid BSE test should not be statistically inferior to that of the currently approved tests.

This document describes the design of such a field trial as developed by the working group that met on 11 – 12 February 2002.

The purpose of this field trial is not to rank the sensitivities of approved and new rapid BSE tests or to find out whether the new rapid BSE tests are able to detect BSE in

[1] As laid down in Annex 10, chapter C to Regulation 999/2001

cattle earlier in the incubation period. These issues should be looked at in separate studies.

3. Estimation of sensitivity relative to approved tests

Sensitivity is the probability that a test recognises confirmed positive test specimens (“true positives”) as positive. Ideally, there should be no “false negative” test results, i.e. the sensitivity is 100%.

It is important that a new rapid test should not perform worse than an approved test. However, absolute equality can only be shown if all available and future samples are tested by both methods. Such an approach is impossible. Therefore, a sample size has to be selected which will demonstrate that, with a given probability, that the sensitivity of the new rapid test when compared to an approved test is not less than a given threshold value. The sample size required also depends on the expected number of true positive results, which would be obtained using the new rapid test (i.e. estimated prevalence of cases in the population). Assuming that a new rapid test might be used in all Member States for a number of years this expected figure might be above 1000.

In order to demonstrate, with a probability of 95%, that the sensitivity of the new rapid test is not below 98% (99%) of the sensitivity of an approved test, the sample size has to comprise at least 138 (258) samples that are positive by one of the approved tests. Samples, which are not detected positive by the approved test are excluded from the study. All of the remaining samples have to be recognised by the new rapid test as positive. Any sample which tests negative with the new rapid test must be re-tested in duplicate from the original sample preparation. Both re-test results must be positive (2 out of 3 have to be positive). The original preparation of the sample must also be re-tested in duplicate with the approved test. Two negative results with the approved test will exclude the sample from the study.

For practical reasons, it was decided that **200 true positive samples** should be tested by a new rapid test compared to approved tests, which would ensure with a 95% probability that the sensitivity of the new rapid test is not below 98,5% compared with the approved test (see annex 1).

The true positive samples can be provided by the National Reference Laboratories (see annex 2). They should be well documented (origin and age of the sample, e.g. sub-population; condition of the sample, e.g. autolysis etc.; brain region used; storage conditions; duration of storage).

In order to avoid test results being discrepant between the new rapid test and the approved test due to an uneven distribution of prion proteins the samples should be made homogeneous. A protocol for the homogenisation treatment which has no or minimal influence on subsequent steps is available[2]. Other equivalent homogenisation protocols can be used if their use is justified to IRMM. A sufficient

[2] The protocol ("Preparation of 1+1 CNS macerates for proficiency testing of BSE-laboratories") may be obtained from Peter Lind, Danish Veterinary Institute, Bülowsvej 27, 1790 Copenhagen V, Denmark, e-mail: PL@vetinst.dk.

number of aliquots representing 500 mg brain stem material should be prepared for this purpose. At least one aliquot must be archived.

The tests will be performed by NRLs of at least two Member States (or one Member State and Switzerland). NRLs will be chosen at the discretion of the company. The maximum proportion of samples tested in a single laboratory in a country must not exceed 70% of all samples.

Derogating from the general rule that sensitivity testing should be carried out in NRLs, certain state owned laboratories may take part in the study if the NRL does not conduct the study itself, if the state owned laboratory is in the possession of a suitable number of positive samples and if the responsible NRL agrees.

Each new rapid test should be compared with at least two approved tests. This comparison has not necessarily to be done in parallel. No more than 70 % of the positive samples should be compared against any one approved test. At least two batches of the test kits should be included. It is the responsibility of the company, which intends to market the new rapid test to select a NRL (or following the derogation a state owned laboratory) as well as the approved tests used for comparison. The company should compensate the laboratories performing the comparisons for all expenses.

IRMM will be notified on the initiation of such studies (where the tests will be carried out? on how many samples? which approved test (s) has been chosen for comparison? and when?) and will collect and evaluate the data. The raw data will be communicated to IRMM at least on a weekly basis. The IRMM will provide a standardised data format.

4. Estimation of specificity relative to approved tests

Specificity is the probability that a test recognises truly negative test specimens as negative. Ideally, there should be no “false positive” test results, i.e. the specificity is 100%.

There is no way to prove exactly that a test is 100% specific. Therefore, a decision has to be made on an acceptable value for specificity. As a specificity of 99.5% would still mean that there are 500 false positive samples in 100,000 tested samples it was decided that the specificity should be between 99.95% (50 false positive in 100,000 tested samples) and 99.99% (10 false positive in 100,000 tested samples). To demonstrate such specificity with a probability of 95% the number of samples to be tested should be between 5,988 and 29,950.

For practicability, it was decided that **10,000 consecutive samples** from healthy slaughtered animals that are tested negative using an approved test should be used for the estimation of specificity in comparison with approved tests (see annex 1).

The tests should be performed with the agreement of the NRL in experienced high throughput routine laboratories. The samples should be prepared according to the company’s protocol. Fresh material immediately adjacent to the usual sampling region can be used for the comparison of the tests. Laboratories of at least two Member States (or one Member State and Switzerland) will be involved at the discretion of the company. The maximum proportion of samples tested in a single laboratory should not exceed 70% of all samples. At least two batches of the test kits should be included.

Each new rapid test should be compared to at least two approved tests. This comparison has not necessarily to be done in parallel. No more than 70 % of the samples should be compared against one approved test.

It is the responsibility of the company, which intends to market the new rapid test to select the laboratories as well as the approved tests used for comparison. The company should compensate the Laboratory performing the comparisons for all expenses.

The NRL will be notified on the initiation of such studies. It will take adequate measures to avoid a disturbance of national statistics.

Discrepant results between the new rapid test and the approved test(s) will be resolved by the responsible NRL and the EU Community Reference Laboratory (CRL) in Weybridge according to an algorithm described below. Confirmed positive results will be excluded from the calculation of specificity.

IRMM will be notified on the initiation of such studies (where the tests will be carried out? on how many samples? which approved test(s) has been chosen for comparison? and when?) and will collect and evaluate the data. The raw data will be communicated to IRMM on a daily basis. The IRMM will provide a standardised data format.

5. Variation in sample quality

In practise, many samples to be tested will be of poor quality (e.g. autolysed or putrified). Therefore, it has to be demonstrated that such conditions do not bias new test performance relative to approved test performance. The positive samples prepared by the NRLs will inevitably contain a proportion of poor quality material. In addition, 200 poor quality negative samples must be studied either by a NRL or a state owned laboratory in comparison with an approved test, e.g. by using material from fallen stock.

IRMM will be notified on the initiation of such studies (where the tests will be carried out? on how many samples? which approved test(s) has been chosen for comparison? and when?) and will collect and evaluate the data. The raw data will be communicated to IRMM on a daily basis. The IRMM will provide a standardised data format.

6. Risk populations

The robustness of a new rapid test should ideally be investigated with routine samples from a high risk population, i.e. a population with a relative high prevalence of BSE. Such a situation arises in some regions amongst fallen stock and emergency slaughtered cattle (e.g. samples submitted to the Newcastle Regional laboratory of the VLA). In addition, these samples are often of the poorest quality of all surveillance specimens. It is therefore highly recommended that companies compare their new rapid test with an approved test on 2000 specimens collected from a high risk population.

7. Resolution of discrepant results from specificity testing (3) and from poor quality samples (4)

Two approaches are used:

a) A homogenised sample preferentially prepared according to the protocol of the Danish Reference Laboratory[3] will be prepared by the responsible NRL from the brain stem which gave rise to the discrepant results. Other equivalent homogenisation protocols can be used if their use is justified to IRMM. Samples of the homogeneous material should be re-tested with the two tests in question by the laboratory where the discrepant results were produced. This approach should be used to resolve discrepant results possibly due to the uneven distribution of the abnormal prion protein in the initial samples.

b) Confirmation will be performed at the CRL according to their established procedures.

8. Description of the test procedure

A number of laboratories will be involved in the evaluation of the new rapid test. It is essential that their results are comparable. It is therefore necessary to have clear, stringent and detailed descriptions of the test procedures. It is the duty of the companies to provide such descriptions and to ensure that they are understood in exactly the same way in all participating laboratories.

After the field trial is finished the laboratories involved should meet in order to discuss whether the written test procedures used during the field trials proved to be accurate and fully understandable. If necessary, test procedures may have to be clarified.

The clarified test procedures or if clarification is not necessary, the test procedures used, will then be defined as part of the approval. Later changes in the test procedure will invalidate the approval.

9. Evaluation of data

The data will be evaluated by the IRMM. The statistical analysis will be designed to demonstrate non-inferiority of the new rapid test to the already approved tests.

[3] see footnote 2.

Annex 1

Statistical background

General procedure for comparing the performance of 2 or more diagnostic tests

Sensitivity and specificity of a diagnostic test can only be expressed in relation to a so-called 'gold standard' (an actual or virtual diagnostic test or procedure, that determines the state of interest with an accuracy of 100%). If such a 'gold standard' is present, the performance of 2 or more 'new' tests can be compared against each other by first comparing each of them against the 'gold standard'. Depending upon the sample size used for this initial comparison, the confidence interval associated with the point estimate for the performance indicator varies. Differences in the proportions of 'correct results' obtained by each test can easily be analysed statistically. If the sample size is large enough even small differences in the actual performance of the two tests can be detected (if they exist) and the two tests can be ranked according to their performance (i.e. it is possible to say which of the two tests is superior). There are also analytical methods available that allow to calculate the sample size needed to demonstrate possible differences in the performance (exceeding a certain pre-set threshold) between two or more tests

If a true 'gold standard' is not available (as it is the case for BSE) other methods have to be used to obtain **estimates** for the sensitivity and specificity of one or more new tests (and actually also for the reference tests used so far).

Methods to obtain these parameters (including an estimate for the prevalence in the population) have been described by different authors (Walter and Irwig, 1988; Hui and Walter, 1980; Gelfand and Smith, 1990, Enoe et al., 2000). These methods either rely on maximum likelihood methods, Bayesian methodology or Gibbs sampling.

The advantages of these new methods are:

- absolute comparison of test performances possible
- new tests are allowed to perform better than established reference test(s).

Unfortunately, these methods also have a long list of disadvantages:

- large confidence intervals
- no general agreement on methodology
- most solutions rely on independence of tests used (which is most certainly not the case with the present BSE-tests)
- no method available to calculate necessary sample sizes
- not suitable to detect small differences in test performance
- two or more populations with distinct differences in prevalence are a prerequisite to conduct a study.

These disadvantages actually prohibit their use for the task of comparing the performance of new BSE-tests to established approved tests, which are used as reference tests.

Therefore, a pragmatic approach has to be chosen.

Approach

A new test has to demonstrate that it is **not inferior** to an already established approved test. As complete equality can not be shown, the term inferiority has to be defined. In our context, a new test will not be inferior, if (with 95% probability) its performance parameters are not below a pre-set threshold (this threshold has to be decided either on the basis of security or economical issues).

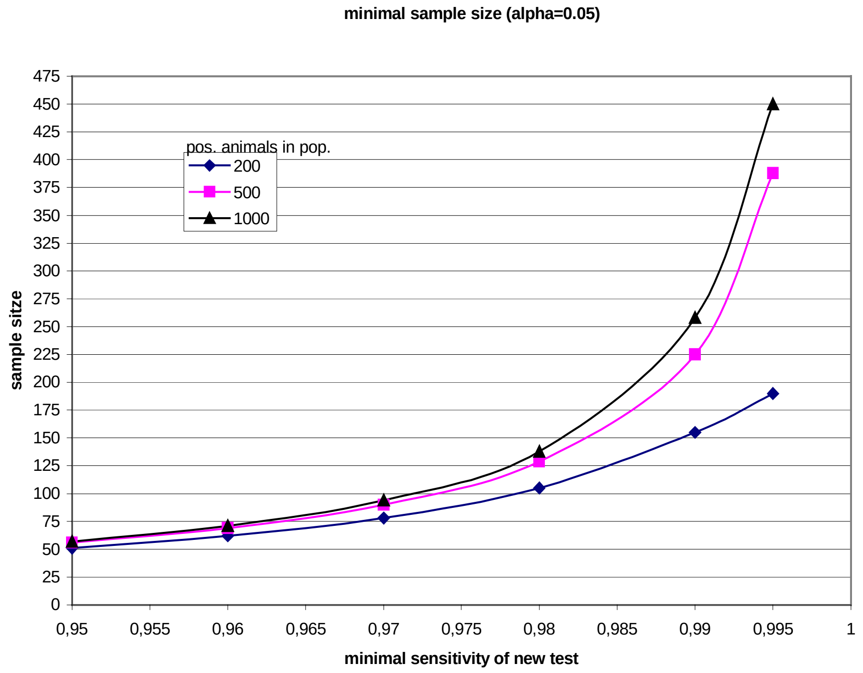
In this context the approved test is treated as 'gold standard' and the performance of the new test is set into relation to this test.

Sensitivity

A new test should (in the view of consumer protection) meet high standards in respect to sensitivity.

In this approach only samples that test definitively positive with the approved test(s) are then subjected to the new test. The table below shows the necessary sample size to demonstrate that the new test meets the pre-set minimal sensitivity criteria (the actual sensitivity of the test might still be better than the one of the approved test but this can not be shown with the chosen study design). No false negative results are permitted for the new test within the given sample size.

Total number of BSE-animals that might be tested with the new test in the future	Sensitivity criteria that the new test must meet minimally	necessary sample size
200	0.95	51
200	0.96	62
200	0.97	78
200	0.98	105
200	0.99	155
200	0.995	190
500	0.95	56
500	0.96	69
500	0.97	90
500	0.98	129
500	0.99	225
500	0.995	388
1000	0.95	57
1000	0.96	71
1000	0.97	94
1000	0.98	138
1000	0.99	258
1000	0.995	450

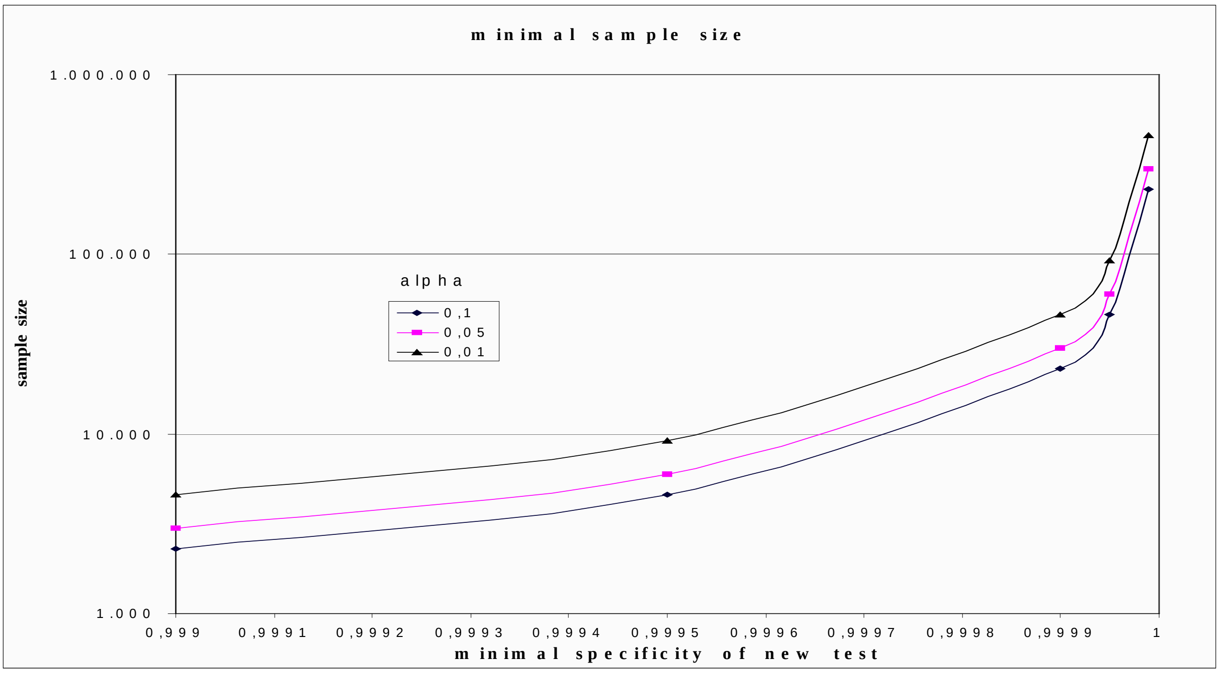


Specificity

A much higher sample size has to be selected to demonstrate that a new test is not inferior to one or more approved tests. Even a minor lack of the new test in specificity might later on lead to an unacceptable number of false positive results under field conditions (due to the large number of - mostly negative - animals tested). This parameter of the new test has no influence on consumer protection but on the practicability of the new test, as all positive results have to be confirmed and in the case of inferior specificity a high proportion would fail this confirmation. Specificity therefore should not be an excluding criteria for a new test, but should be evaluated and clearly stated.

The table below shows the sample size necessary to demonstrate that a new test meets at least a pre-set specificity (its actual specificity might be well above this value).

Minimal specificity of new test	Necessary sample size	alpha	Minimal specificity of new test	Necessary sample size	alpha
0.999	2301	0.1	0.999	4600	0.01
0.9995	4606	0.1	0.9995	9200	0.01
0.9999	23025	0.1	0.9999	46000	0.01
0.99995	46050	0.1	0.99995	92000	0.01
0.99999	230200	0.1	0.99999	460800	0.01
0.999	2994	0.05			
0.9995	5988	0.05			
0.9999	29950	0.05			
0.99995	59900	0.05			
0.99999	299600	0.05			



Possible problems with this approach:

A new test might well be more sensitive than the confirmation test. This could lead to a misjudging of the apparent specificity of a new test, as in fact positive animals (that can not be confirmed) are attributed as being false positive. To circumvent this problem, this trial should ideally be conducted in a population that is either free of BSE or has a low prevalence (for instance healthy slaughtered animals).

Reference:

Enoe C., Georgiadis M. P., Johnson W. O., 2000. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev. Med. Vet. 45, 61-81.

Gelfand, A., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. J. Am. Statist. Assoc. 85, 398-409.

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36(1), 167-171.

Walter, S.D., Irwig, L.M., 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J. Clin. Epidemiol. 41(9), 923-937.

Annex 2

Approximate number of positive samples available from different National Reference Laboratories

CH	80
DE	70
UK	50
IE	26
PT	16
IT	16
ES	16
FR	?[4]
BE	35
NL	?[4]

[4] At the time of adoption of the opinion the numbers were not available.