EVALUATION OF ANALYTICAL METHODS WITHIN A CONTEXT OF USE

O'RANGERS J.J.
BERKOWITZ D.B.

Food and Drug Administration, Rockville, Maryland

Since 1985, there has been a growth of interest in the private and government sectors, in the application of screening tests for the detection of animal drug residues. Although screening tests can be based on any aspect of analytical technology, most of the screening tests being commercially produced for animal drug residues are based on immunoassay or biological receptor technologies. The commercial products are manufactured in a "test kit" format which are intended to be a self contained, complete analytical test for animal drug residues similar to test kits that are marketed for human diagnostic purposes. This paper is about the evaluation of screening methods. The main point is that methods can not be evaluated in isolation. There is no such thing as a "good method" or a "bad method" because methods must be evaluated *for a particular purpose.* A method which is ideal for one purpose may be totally inadequate for another. Thus, to ask if a method is good is to ask only half of a question; one must ask if it is good *for a particular purpose.* Methods must be evaluated within a context of use.

Screening tests provide several exciting capabilities to animal drug residue detection. For example, because the tests do not require complicated instrumentation, they are usually very rapid in performance with analytical results being achieved in minutes. Not only can more tests be performed in a given time period, but many screening tests can be used outside the laboratory. This capability is an advantage in residue control and public health protection, since an initial analysis is a real possibility at the level of drug use. The use of screening tests in residue detection can be exemplified by the tests that are commercially available for drug testing in milk and their application in milk testing. The acceptance of animal drug screening tests for regulatory or public health uses should be based on adequate fulfillment of the following three considerations.

THE SCREENING TEST MUST BE AVAILABLE FOR USE

The use of recommendation of a screening test by public health or regulatory agencies must be practical. Often tests and methods are described in the scientific literature that would appear to offer good potential for screening test application. However, the general accessibility of these methods is usually very limited unless the reagents or tests can be purchased commercially or will be provided by some other means, such as government contract or drug sponsor. It is important to keep in mind that research is only the first step in the development of useful tests.

A mature, fully characterized and practical screening test that is suitable for routine regulatory or public health use requires additional work to ready the test for real world application.

THE PERFORMANCE OF THE TEST MUST BE ADEQUATELY CHARACTERIZED

The test must be demonstrated to perform with acceptable figures of merit for accuracy, specificity and reproducibility. Time will not allow a detailed discussion of all the technical aspects of screening test performance evaluation, however, there are no esoteric performance requirements that are peculiar to screening tests. The performance elements for screening tests are similar to the performance elements that are assessed for other analytical methods.

ALL SCREENING TESTS SHOULD HAVE CONFIRMATORY METHODS

While screening tests provide presumptive information on the existence of a drug residue in a test sample, they do not provide definitive information on the identity of specific drug residues. Often, screening tests are designed to test for multiple residues of various drugs. An example of such a test is the Charm II receptor assay for beta-lactams inmilk. This test can give an initial alert that a beta-lactam residue may be present in a milk sample. However, the determination of the specific beta-lactam drug that may be present in the sample requires verification by more specific analytical techniques.

Furthermore, many screening tests for drugs in milk that are currently on the market give a "yes" or "no" answer, i.e., they are designed to provide an indication of the presence of a drug residue in milk above a defined quantitative level. These tests do not give a quantitative assessment of the residue that may be present. In many cases, it is important to know how much of a given residue may be present in order to take appropriate action. The required confirmatory or quantitative information can be provided by a suitable method.

Good methods frequently evolve in a context of use, and are developed for a particular purpose. The USDA STOP TEST was designed for the rapid determination of antibiotics in meat. A sterile cotton swab is exposed to tissue fluids and placed on a petri plate seeded with bacteria sensitive to several antibiotics. If the animal fluids contain antibiotics, a zone of inhibition is seen around the cotton swab. The test is based on well-known principles, and is simple, inexpensive, portable, and sensitive. As a screening test it is excellent, but as a confirmatory test legal action it would be totally unsuitable, because the cause of the inhibition is not identified; it could be from an antibiotic residue or from an elevated serum component of the animal. The suitability of the test must rest on the context of its use.

A screening test can be simply defined as a test that gives a reliable indication that the analyte(s) of interest are not present in the sample at hazardous or violative levels (O'Rangers, J., 1990). This requires that the screening test be developed with a limit of detection (LOD) at a level that will give a high degree of confidence that the tolerance or violative levelwill be detected. In designing a screening test, this usually means that the LOD of the test will be optimized below the hazardous or violative level so that the higher hazardous level will have a high probability of causing a positive result, or in the jargon of test kit users, "lighting up the test".

This technical specification also means that the screening test can give positive results below the tolerance or violative level. The frequency of positive results will be somewhat smaller than at the tolerance level, but can, and probably will occur. Without further quantitative analysis or comparison to standards, it is difficult to be sure whether a positive screening test result is actually at the tolerance or violative level. This is a technical cost that is incurred in developing screening tests to monitor hazardous or violative levels or animal drugs and a major technical objective in the development of these tests is minimize the band of uncertainty for the screening test.

The great value of screening tests is the degree to which they reliably indicate samples for which there is no regulatory or public health concern so that the commodity can enter the food supply. A positive result indicates that there is reason to withhold the food commodity and that follow-up action is needed.

It is apparent that the suitability of a method for a particular objective may be determined by factors external to the test itself. If the samples containing the analyte are well separated from the background or cross-reacting substances, the number of false positives will be minimized. However, if the background level of the analyte or of some cross-reacting material is sufficiently close to the region of the analytical response given by the analyte, there would be a considerable amount of overlap in the analytical response regions of both the analyte and interfering materials and the percentage of false positive determinations may be unacceptable. For example, the use of the diazotization reaction for the determinationof aromatic amines to determine sulfamethazine may be perfectly suitable for drug samples, but in animal tissues the background level of aromatic amines is very high, so one must first separate the sulfamethazine from the cross-reacting material. The test is useful when the background is low, but not useful when the background is high.

Furthermore, screening tests can be based on any type of technology and may involve considerable technical complexity. Multi-residue chromatographic methods designed to separate and quantitate multiple drugs or pesticides require experienced laboratories and analysts to be used effectively.

These types of screening tests are intended to provide efficient economies of scale for the expert laboratory so that analytical coverage can be maximized. These types of screening methods will not be discussed.

Rather the types of screening tests that are of interest have the following characteristics:

1. They usually do not depend on complicated analytical instrumentation. Analytical signals are usually generated colorimetrically and use simple detector systems.

2. Although the tests do not depend on complicated instrumentation, they do generally depend on reagents that have complex scientific principles of operation. The quality of these reagents is usually dependent on the processes use to produce the reagents.

3. They generally do not require lengthy, multi-step sample extraction procedures although some minimal sample preparation may be required.

4. They usually determine multiple analytes although they can be designed to detect one analyte with relatively high specificity.

Screening tests having all or some of the above characteristics are usually ligand assays such as immunoassay or receptor assays and are typically provided in a test kit format. The manufacture of reagents into a test kit involves the configuring of reagents to achieve the desired performance characteristics for the test. For example, the sensitivity (limit of detection) of immunoassays partially depends on the concentration of the antibody or other receptor used in the assay. Also the composition and components of the associated reagents such as buffers and substrates determine the performance of the test. The important point is that the performance characteristics of any screening test kit depends on the unique interaction between the specific components of the test kit. Any changes in the individual test kit components usually can result in altered performance of the test kit.

When selecting a test kit for use or evaluation, the user may consult with the test kit manufacturer for the following information:

1. The test kit should have specific performance specifications set for the over-all test kit. This is the responsibility of the test kit developer or manufacturer. The test kit performance should be confirmed by the manufacturer by the use of alternative methods or definitive confirmatory techniques.

2. If individual reagents are separately purchased to be used in a screening test procedure developed by the user, each individual reagent should have performance specifications established by the manufacturer so that the user can continue to select reagents that meet consistent performance specifications.

3. Evaluation of test kits should be done on the test kit as a whole. This should have been performed by the manufacturer and should be verified by the user in an independent validation study. Alteration of any test kit component as part of the evaluation procedure may invalidate the test as an artefact of testing. Test kit evaluation protocols should observe the response of the test kit to appropriate physical and chemical tests. The evaluation protocol should not significantly change the characteristics of the test kit components as provided by the manufacturer.

4. The test samples used in the evaluation of the test kits should emulate as closely as possible the types of samples that are likely to be encountered by users of the test kit. This means that in addition to using fortified test samples, test samples that contain naturally incurred analyte(s) should also be used when possible.

If incurred samples are not routinely available, well designed field tests could be used to fulfill this recommendation. In the case of field tests, the results of the test kit should be compared to the results achieved by the accepted method of analysis for the tested analyte.

Using the information provided by the test kit manufacturer, the test kit user should develop a validation plan for all test kits that are being considered for use. The purpose of the validation plan is to assure that the test kit performs acceptable in the user's application and the data developed by the screening test is useful to the user to support the intended regulatory or public health action.

All methods, including screening tests, have innate characteristics, the two most frequently discussed are sensitivity and specificity. These characteristics are very helpful in judging the usefulness of an assay. However, sensitivity and specificity also vary with the intended use. Some examples are discussed below.

SENSITIVITY

There are some differences in the way sensitivity is described.
For example, sensitivity and the limit of detection are used in different ways. A method is frequently referred to as sensitive if it can detect a low level of analyte. Chemists refer to the lowest level of detection as the limit of detection, and method sensitivity as the increment in response relative to the increment in concentration. An example will illustrate the different usages. Assume that we have two methods, method A and method C. Furthermore, assume that method C has a lower limit of detection than method A, but the response curve for

method A has a greater slope than method C. The typical conclusion is that method A is more sensitive because the analytical response is greater for each increment in concentration. However, limit of detection and sensitivity are not entirely independent. For example, method C has a lower limit of detection because it has a lower background. If the two methods had the same background, method A would have the lower limit of detection. This is because the point at which the sample response becomes significantly different from the background, i.e., the concentration at which the background and sample signals become significantly different, is lower in A than in C. To calculate this concentration, we take the simplest case and assume the standard deviations are equal at zero concentration and at the limit of detection. The variance of the difference is twice the variance, so the estimated standard deviation of the difference is $S = (2)^{½}S_O$ where S is the standard deviation of the difference, and S_O is the standard deviation of the blank or sample, because they are assumed equal. At the 95 % confidence level, the signal becomes significant when it is two standard deviations greater than the blank signal, and $2S = 2(2)^{½}S_O$. The limit of detection for method A is lower *because* the slope is grater; i.e., the signal becomes significantly different from the blank signal at a lower concentration, Thus, the improved sensitivity of method A reduces the limit of detection.

Many screening tests are dichotomous "yes" or "no" tests. If this is the case, and the test is designed to indicate the presence or absence of an analyte, the limit of detection must be known, so that the lower concentration limit of what will be detected is known. If the purpose of the assay is to establish that the analyte does not exceed some established level, the limit of detection is far less important, and the reliability of the test must be known across the level of interest. This topic is discussed in more detail below.

For immunoassay, obtaining this kind of information is somewhat more difficult because the antibody titration curve is sigmoidal rather than linear. Nonetheless, useful information about the characteristics of a screening test can be derived from the sigmoidal curve which is also known as "characteristic operating curve". In this technique, a panel of test samples is produced by fortifying control matrix with the drug of interest.

The test is run on each sample at its respective concentration level. For good statistical confidence in the measurements, it is recommended that 15 to 20 replicates be run at each concentration level. For rapid screening tests this should not represent an undue analytical burden.

The results are plotted as the percent of samples that are positive at each concentration level. The data from this curve will indicate the following:

1. The concentrations of drug residue that can be detected and the confidence associated with each level.
2. The false positive samples that will be expected.
3. The ability of the test to discriminate between sample residue concentrations.
4. If the test is run over a given time period, the stability of the test can also be evaluated.

The operating characteristic approach is not new. It is an example of an adaptation of the Probit concept elaborated by a number of research workers notably Gaddum, Trevan, Bliss and Finney. In addition, the four parameter logistic algorithm developed by Rodbard, *et al.* could also be adapted to analytically describe the operating characteristic curve of screening tests.

A note of caution is needed at this point. The degree of confidence that can be assigned to a result at any concentration level, is partly a function of the number of replicates assayed at each concentration level. While the use of 15 to 20 replicated will give a good assessment of the best confidence that can be assigned to each assay will depend on the number of replicates actually run in practice. This is a critical point to keep in mind when deciding how a screening test is to be used. If the test is to be used for quantitative purposes, a high degree of confidence in the analytical results is usually required. This requirement could require sample replications that are impractical and uneconomic for routine screening test application for quantitative purposes.

These difficulties are circumvented by use of the logit transformation which linearizes the curve.

In general, an estimate of the precision at any concentrations level can be estimated from the empirical relationship derived by Horowitz. The equation for relative standard deviation is: RSD(%) = $2^{1-0.5\log C}$, where C is the concentration of interest. For example, when the analyte is pure, C is 1, and the RSD should not exceed 2 %. If the analyte is at 1 ppm, C = 10^{-6}, and the RSD may be as high as 16 %. This is a useful empirical equation, with some theoretical rationale (AOAC and Anal. Chem.).

SPECIFICITY AND INTERFERENCES

The determination of assay specificity involves the evaluation of the extent to which the assay reacts only with the compound of interest. These studies are also known as cross-reactivity studies. They are performed by assessing the reactivity of the test with structural variants of the test molecule or related chemical substances that may also be present in the sample.

Specificity can be exquisite in immunoassay, and at the same time, it can be exquisitely vexing. For example, an immunoassay for the penicilloyl group is very sensitive and is able to detect the penicilloyl group at very low levels (Humphries). However, when the assay was used to monitor the pharmacokinetics of penicillin elimination from the serum of treated animals, the situation became complicated. Although the antibacterial activity was all eliminated from bovine serum within 24 hours after injection, the serum levels measured by immunoassay remained high for at least several weeks. This was because the immunoassay measured not only the free drug, but also the penicilloyl groups bound to protein in the serum. The half-life of penicilloyl groups covalently bound to serum proteins are roughly equal to the half-life of the proteins in the circulation. Again, the intended use determines the suitability of the assay.

The definitions of sensitivity and specificity in the clinical literature are particularly useful, because they describe how a test performs in particular situations. In the clinical literature, sensitivity means the percentage of true positives detected as positive. Specificity is defined as the percentage of "negatives" reported as negative. Thus, the ideal test is *both* 200 % sensitive and 100 % specific. In some ways, an even more revealing measure is called the predictive value of a test.
The predictive value of a positive test is plotted against the pretest

probability of a sample beingpositive. Let's assume that we get 10 % false positives and 10 % false negatives, and that the true percentage of positives in the population is 1 %. If 1000 samples are measured, the pool should contain 10 positives (i.e. 1 %), but only 9 are detected because one is a false negative. On the other hand, we have 990 true negatives, but of these, 10 % or 99 are false positives.

We report 100 positives when, in fact, we have only 10. So, the predictive value of a positive test is only 10 %. But we have 990 true negatives and report 990 - 99 = 891, so the predictive value for negatives is about 90 %. These estimates depend highly on the true frequencies in the population. Using the same false positive and false negative rates of 10 %, but increasing the number of true positives to 10 %, makes the test more useful. In the population of 1000 we now have 100 true positives. We report 90. Thus, the predictive value of a positive test is 90 %, a far better performance than was the case when the true frequency was lower.

Our certainty has been increased by a factor of 9 as a result of the situation in the population, i.e., a circumstance outside of the characteristics of the test itself. This is a dramatic demonstration of the influence of the intended use on evaluating test suitability.

Interference, can be practically assessed by using the test under field conditions, using samples containing known quantities of test analyte. The test should provide accurate and reproducible results in the real world environment. Any difference in the performance of the test between the laboratory evaluation and the field test should be resolved before the test is incorporated into a testing program.

ASSESSMENT OF METHOD PERFORMANCE USING ACTUAL TEST SAMPLES

The following sets of test samples should be used in the evaluation of the test:

Set 1: Background or blank samples should come from animals or control. The data from these samples are used to determine the false positive ratio which is used to establish the "diagnostic" specificity of the test.

Set 2: Samples that are *known* to contain a definite quantity of the test analyte. Typically, these samples will be generated by "fortifying" or "spiking" a suitable matrix with the test analyte. The matrix could be a negative control or may be a sample containing potential cross-reacting or interfering substances. The data from these samples are used to determine the false negative ratio and to establish the "diagnostic" sensitivity of the test.

Set 3: Samples that contain residues of the test analyte that have been naturally incurred in samples from animals or other test article that have been exposed to the test analyte. Different cohorts could be examined in this experiment. For example, normal animals containing the test analyte would be tested followed by examination of cohorts of animals with pathologies or containing substances that are likely to occur and be associated with use of the test analyte. It should be important to understand the response characteristics of the screening test to these cohorts.

Ideally, these test samples should be maintained in a test panel that can be used to evaluate different lots of test kits or reagents. If a test panel is maintained, the stability of the analyte in the test matrix must first be determined.

CONFIRMATORY ANALYSIS

There is no single correct procedure or strategy for confirming animal drug residues in meat, milk or eggs. The procedures or the extent of analysis that is needed for confirmation depends entirely on the intended use of the analytical results.

For example, a high degree of certainty is required in establishing the identity and quantity of a drug residue in meat, milk or eggs if the intent is to assess penalties against individuals or organizations for violation of laws or regulations. The very best scientific procedures should be used in these cases not only out of a sense of fairness and official responsibility, but ensure the capability to pursue future regulatory cases.

However, in public health monitoring, definitive identification of specific drug entities is desirable but not strictly necessary in order to take effective action. It is necessary that the analytical results indicate that there is a high probability that a food safety problem may exist in the sample and that further action is warranted to determine the disposition of the food commodity.

In some cases, confirmation may not be needed at all. For example, in establishing quality assurance programs for the acceptance if raw materials for manufacturing, specifications can be established that require the raw material to pass a specific test or battery of tests. The only requirement in this case is that the tests used be well characterized and validated for the proposed use.

There are several approaches that can be used to address confirmation of results:

a. Use a definitive reference method to confirm the initial analysis

In regulatory analysis the method of choice is mass spectrometry. Mass spectrometry gives specific information on the identity and structure of the compound of interest. Coupled with techniques such as gas chromatography, this becomes a very powerful confirmatory tool for both quantitative and qualitative assessment of chemical residues in food. Heat labile chemicals can also be confirmed by interfacing HPLC with mass spectrometry, such as HPLC-Thermospray mass spectrometry.

b. Confirm with several methods or tests

Ideally, the tests used should assess different chemical characteristics of the analyte. The chromatographic behaviour of the analyte under different conditions can be effectively used. Normal phase, reversed phase, size exclusion and ion exchange are all examples of chromatographic conditions that operate on different physico-chemical principles and when used with appropriate standards, can give a more definitive insight into the identity of a test analyte.

Different detectors can also be used to exploit the different chemical features that may be characteristic of a given analyte. Photodiode array, ultraviolet, fluorescence, and electrochemical detection are all commonly used in residue analysis and can be used in an on-line mode with chromatographic systems. Gentian violet is an example of a compound that can be determined using ultraviolet or electrochemical detection.

The coupling of chromatographic procedures with immunochemical techniques can provide a very sensitive and specific method for either determinative or confirmatory analysis. Immunoaffinity chromatography has been used extensively in protein chemistry research and is finding increasing application in animal drug residue analysis. Immunoaffinitychromatography involves the coupling of antibodies to a chromatographic support, thereby producing a relatively specific chromatographic media for the drug(s) of interest.

The utility of immunoaffinity chromatography in animal drug residue analysis is persuasively shown in the Proceedings of the EC-workshop on The Use of Immunoaffinity Chromatography in Multiresidue and Conformation Analysis of Beta-agonists in biological samples (Van Ginkel, L.A., et al., 1991).

Immunoaffinity chromatography has been used successfully for the removal and concentration of aflatoxin B1, B2, G1 and G2 allowing the detection of as little as 0.5 ng of aflatoxin (Candlish, A.A.G., 1988).

Immunochemical reagents can also be used as an off-line chromatographic detector where fractions of the chromatographic eluate can be assessed by either RIA or ELISA for the analyte of interest. If the antibody used is very specific for the analyte of interest and the antibody reactivity is known to be sensitive to small variations in the structure of the analyte tested, positive reactions with the method are strongly indicative. That and the analyte of defined structural characteristics are likely present in the sample. Full rigorous confirmation, of course, would depend on further analysis by mass spectrometry. I have personally used this approach in the analysis of diethylstilbestrol in bovine liver.

If the antibody used as a detector is not specific for the analyte of interest, the chromatographic conditions can be adjusted to optimally separate the analyte from other interfering or cross-reacting components that may be in the sample. Thus, the power and flexibility of the linkage of chromatography with immunochemistry is readily apparent.

Useful screening technology using solid-phase techniques is not limited to immunoaffinity procedures. A new variation of solid-phase extraction has been developed which permits the very rapid extraction of drugs and other chemicals from complex biological matrices. This procedure is known as Matrix Solid Phase Dispersion (MSPD) and has been developed with the support of the US Food and Drug Administration, in the laboratories of Professor Steven A. Barker at the Louisiana State University, School of Veterinary Medicine, Baton Rouge, Louisana. This technique has been used for the separation of a variety of commonly used antibiotics from animal tissues (Barker, S.A., Long, A.R., 1992). The use of this technique holds great promise for a simple clean-up procedure for immunoassay screening methods.

c. Define an existing method as a standard or reference and compare the new or proposed test to the standard

This is a practical strategy which makes a great deal of sense when the standard method has been well characterized and demonstrated to be reliable.

It is quite likely that a reliable body of data on the analyte of interest already has been developed through use of the standard method.

This data will be a valuable source of historical control values for use in the evaluation of the new method.

A key point to keep in mind, is that the comparison of the new method with the standard method must be done on a standardized procedure for evaluating the data that each method generates. For example, any correction factors that are used, such as recovery corrections, must be normalized.

References

Barker, S.A. and Long, A.R. J. Liq. Chromatog, 15, 2071, 1992.

Candlish, A.A.G. Intl. J. Food Sci. Tech., 23, 479, 1988.

O'Rangers, J. in Immunochemical Methods for Environmental Analysis, pp.27-37, VanEmon, J. and Mumma, R. (eds), 1990.

Van Ginkel, L.A., van Rossum, H.J. and Stephany, R.W. (eds). Commission of the European Communities, RIVM, Bilthoven, The Netherlands, 1991.