Conference on Applied Statistics in Agriculture 2006 - 18th Annual Conference Proceedings

DECISION QUALITY METRICS – A TOOL FOR MANAGING QUALITY OF REPEATED BIOASSAYS

Nancy Ferry

William Letsinger

Decision Quality Metrics – A Tool for Managing Quality of Repeated Bioassays

Nancy Ferry[1], William Letsinger[2]

[1]DuPont Crop Protection Products, Test Quality and Statistics, Newark, DE 19714, USA
[2] E. I. du Pont de Nemours and Company, Wilmington, DE 19898 USA

Abstract

Bioassays are often used in tiered screening systems to detect potential products, such as crop protection products. Often these assays are not replicated. The ultimate products of these bioassays are decisions, with biologically “active” compounds advanced to the next level of screening. Activity is determined by the response of the test organisms (e.g., weeds, insects or fungi) to each compound. The reproducibility of the bioassay is crucial. There are two types of possible errors in screening, false positives and false negatives. The quality of the decisions based upon these bioassays can be monitored through time using controls. This paper will discuss Decision Quality Metrics, quality control metrics customized for bioassays used to select the most “active” compound. These metrics monitor the reproducibility of the screens, translating bioassays responses to controls into potential impact on decision making.

Keywords: Bioassays, high throughput screening, quality management, decision quality

1. Introduction

Companies discover, develop and sell crop protection products to control agricultural pests (e.g. weeds, insects, plant disease). Companies may test tens of thousands of compounds each year looking for new products. Compounds may be selected for testing based on chemical and physical properties that are believed to predispose a compound for biological activity on agricultural pests. These compounds are often sent through a tiered screening system, which uses bioassays as detectors of biological activity.

New crop protection products are difficult to find. A successful product candidate must control a spectrum of agricultural pests, be safe to the crop, safe to agricultural workers and have safe levels of residue on food crops. It must also be safe to beneficial insects, nontarget plants and other nontarget organisms. It must be safe to the ground water, surface water and living organisms in these environments (e.g. aquatic invertebrates, fish). One implication of these challenges is the use of high throughput product screening by crop protection companies testing large numbers of compounds.

High throughput screens that require only small quantities of each compound are used to assess the biological activity of each compound on target pests. Scientific expertise is used to miniaturize the detector bioassays, especially in early screens. Increased automation makes screening large number of compounds feasible. One drawback of this approach is the availability of little or no replication for experimentals. Some screens may use 96 well microtitre plates, where each well is a test unit for an experimental compound. An advantage of automation is that it facilitates the introduction of increased replication for positive and negative controls which are the basis for the generation of Decision Quality Metrics.

Bioassays are designed to detect herbicidal, fungicidal or insecticidal activity. A bioassay consists of a target pest organism (e.g., insect, weed or plant disease pathogen), growth media (e.g., soil or agar) and often a host plant, all in a test unit. The experimental compound is applied to a test unit, which is incubated under prescribed conditions and, after a specified duration, the response of the test organism to each compound is measured. The bioassays have been designed by scientists to optimize sensitivity, response detection, reproducibility, robustness and low compound consumption.

A screen may be made up of multiple bioassays, each for a different target pest organism. Figure 1 is a graphical representation of the way experimental compounds might move through a tiered screening system. A compound advances to the next level of screening after it demonstrates biological activity over a specified threshold.

The goal of a tiered screening system is to eliminate inactive compounds from further consideration and to detect active compounds. The end products of each screen are decisions, to advance or not to advance to the next tier, for each experimental compound. The high throughput screening model has made screening, especially the early level screening, much like a production line process. The disciplined processes of quality management (e.g., procedural control, change management, audits and metrics) may be implemented to assure that every experimental compound is treated in the same manner. Customized quality management control charts may be used to monitor repeated bioassays. Customized metrics, called Decision Quality Metrics, will be discussed in this context, but the methods apply to any repeated bioassays.

2. Quality Management for Repeated Bioassays

The components of quality management (QM) that can be applied to a screening process are (1) Standard Operating Procedures (2) Change Management (3) Audits, and (4) Metrics. The application of quality management techniques to repeated bioassays has the goal of every compound experiencing the same experimental conditions.

There are many opportunities for unwanted variation (i.e., experimental error or noise) to enter the bioassay-based screening process. With each run, target pest organisms are prepared (e.g., solutions of plant pathogens) or selected (e.g. plants) for use in screening. Compound samples are physically weighed, put into solution, and applied by pipette or sprayer to the test units. Then each test unit is incubated in prescribed conditions for a specified duration. After incubation, each compound is evaluated for efficacy of pest control. Each compound is in a test unit. The unit may be a pot with multiple weeds or a single well of a 96 well plate containing a target pest organism. Ideally all these procedures are carried out uniformly, so that each test compound experiences the same handling, growth conditions and evaluation process. The discipline of quality management can help reduce experimental error.

Promoting the uniformity and discipline of quality management to research and development researchers poses some challenges. Generally, researchers are trained to make discoveries using the scientific method, where each run of an experiment is fine tuned, based on the learnings from the prior experiment. The change management component of QM is useful in reconciling the iterative tweaking of the scientific method and the uniformity of execution valued by quality management.

Standard operating procedures (SOPs) can be put in place to insure that bioassays are conducted as designed. Change management is a disciplined method of introducing improvements into a process. Improvements are optimized outside the routine screening system and then introduced into the screening system only after procedural controls are put in place. Using change management, researchers employ their creativity to improve assays outside of the routine screening process. Once an assay format is in finalized, it is run as reproducibly as possible. Routine audits assess the effectiveness of the quality management program.

One metric used for a screening process is the percent of compounds tested that show biological activity. This metric, dependent on the chemistry selected for testing, is a good measure for the effectiveness of the compound selection process, but not a good measure for the consistency of the bioassay process. The percent of compounds that are active is valuable feedback to those who select which compounds should be tested. The physical and chemical properties of the ‘active compounds’ can then be contrasted with properties of ‘inactive compounds’ and provide guidance on future compound selection criteria. A useful metric for assessing the consistency of the bioassay process should track and quantify experimental error.

Conventional run charts for the negative controls (i.e., test units without any compound applied) are used to monitor the stability of the detector system (Farnum, 1994). For example, p charts can be used to track the percent mortality of untreated test units in an insect screen. Selected rates of commercial crop protection products are used as positive controls. Proposed in this paper is a metric, based on positive controls, that translates variability in the bioassay process into its potential impact on decision making.

3. A Possible Metric - Z-Factor

One statistical parameter proposed by Zhang, Chung and Oldenburg (1999) for use in evaluation of high throughput screening assays is called the z-factor. The z-factor uses the signal window, or separation band between positive and negative controls to evaluate assay quality. The z-factor expresses the separation band as a percentage of the dynamic range, which is defined as the distance between averages of positive controls (C+) and negative controls (C-).

The formula is:

$$\text{z factor} = \frac{\left|\mu_{c-} - \mu_{c+}\right| - \left(3\sigma_{c-} + 3\sigma_{c+}\right)}{\left|\mu_{c-} - \mu_{c+}\right|}$$

$$\text{z factor} = \frac{\text{separation band}}{\text{dynamic range}} = \frac{\text{dynamic range} - \text{data variation bands}}{\text{dynamic range}} \qquad (1)$$

Figure 2 is a graphical display of the z-factor calculation. This metric is based on the signal to noise ratio of the difference between the minimum and maximum signals, assuming normality for both the positive and negative controls. The authors describe an assay with a z factor value between 0.5 and 1 (i.e., more than half of the dynamic range is a separation band) as an excellent assay. The z factor has the advantage of being a dimensionless simple statistic. This metric addresses the signal to noise ratio, but does not take into account the mean to variance relationship in bioassays. Bioassays have a distinctive mean to variance relationship, where the variability is greater around 50% control responses and much less near both the 0% and 100% control responses. See Figure 5 for graphical representation. An alternative approach is the operating characteristic curve, which works directly with probabilities of a compound being passed as active, given the mean activity of that compound.

4. Operating Characteristic Curves and Error Rates

In bioassays used to detect compounds with biological activity, each compound is assessed as active or inactive based on the observed level of biological activity. There are two possible correct decisions, correctly identifying an active (i.e., sensitivity) and correctly identifying an inactive (i.e., specificity). There are two possible incorrect decisions. A false positive occurs when an intrinsically inactive compound is identified as active. A false negative occurs when an intrinsically active compound is identified as inactive. These two errors have different ramifications. The cost of a false positive is the time and resources used in the next tier of screening to bioassay the compound. The cost of a false negative is difficult to quantify. What if you missed the next blockbuster product?

An operating characteristics curve describes the reliability of detecting the true, but unknown, level of activity. (Farnum, 1994) Figure 3 shows an ideal operating characteristic curve for a screen with a goal of advancing every compound with activity of 80% control or greater. The true, but unknown, activity is estimated by the mean response across replicates and time. The mean activity is shown on the x axis and the probability of advancing this compound to the next tier is on the y axis. Figure 4 shows a more realistic operating characteristic curve. The area under the operating characteristic curve, to the left of the x value of 80%, represents the false positive rate. The area above the operating characteristic curve, to the right of the x value of 80% represents the false negative rate. This equating of area with error rate is true only under the assumption that compounds are randomly selected from a group of compounds whose level of activity follows a uniform distribution.

Distinct patterns of variability exist in bioassays and depend on the level of stimulus, background response and sensitivity changes in the target pest organism. Figure 5 shows a hypothetical graphic demonstrating the impact of biological variability on error rates. In this graph, the rate tested is on the x axis and the percent control is on the y axis. Bioassays of compounds demonstrating intrinsic activity show a distribution of responses that follow a dose response curve. The dose response curve goes through the mean response for each rate tested. Replication, both within and across runs, at each rate tested would yield a distribution of responses. False positives are that part of a distribution of possible responses for a rate of a truly inactive compound that are observed with activity greater than the promotion criteria. False negatives are those realizations of a rate of a truly active compound that are under the promotion criteria.

If there is only a small quantity of each experimental compound available for testing, it may be difficult to confidently assess its 'true' or intrinsic activity. The next section will describe how positive controls, commercial crop protection products, can be used to generate an operating characteristic curve and metrics.

5. Use of Standards for Metrics

To best monitor the consistency of a repeated bioassay, a stimulus is needed that can be applied and replicated run after run to the bioassay. Selected rates of commercial crop protection products, which will be referred to as standards, can be used to supply these reproducible, repeated stimuli.

Rates of each standard should be selected to elicit a dose response (i.e., 20% control to 80% control of the target pest organism). Rates of standards can be used to mimic experimental compounds of various activity levels. A reproducible stimulus to a bioassay can be obtained by subjecting test systems to a fixed rate of a commercially available 'standard' crop protection product for the assay pest. This testing of standards allows the characteristic variance to mean relationship to be generated. A rate of a standard that is much lower than the recommended use rate may cause a response in a test organism, similar to that which may be caused by an experimental compound with weak activity. It is assumed that standards will go through the same bioassay process as experimental compounds. The responses generated when the same stimulus is repeatedly applied to a bioassay allow for the quantification of variation at each level of response and the resulting error rates.

One assumption underlying the use of standards to generate metrics is that the variability in a bioassay is more dependent on a bioassay's response to a level of stimuli than it is on the nature of the stimulus itself. For example, the distribution of responses from a high dose of an intrinsically weak compound that gives a mean response of 50% control is the same as the distribution of responses of a low dose of a very potent compound that gives a mean response of 50% control. Our experience, having examined thousands of dose response curves, supports this assumption. The Decision Quality Metrics translate this variability into potential impact on decision quality, by calculating estimated false positive and false negative rates.

In the calculations of these estimated error rates, the promotion criteria used for experimental compounds is applied to the observed responses of the standards, thus treating the standards results as if they were experimental compounds. Each screen has promotion criteria that must be met for an experimental compound to go to the next level of screening (e.g., if observed mean response is greater than 80% control of pest). Standard test units showing biological activity over the promotion criteria are counted as actives and units showing activity under the threshold are counted as inactives. These hypothetical "advancement decisions" are then compared to the intrinsic activity for the standards, which can be estimated by the mean response across runs.

6. Assay and Screen Level Metrics

Screens are made up of multiple assays, targeting different pest organisms within the discipline. Promotion decisions are made for each experimental compound. It is possible that activity of an experimental compound on multiple assays is required for promotion. We calculate both assay level and screen level metrics. Assay level metrics are informative to screen operators, monitoring reproducibility at the assay level. The assay level metrics are sensitive to assay specific fluctuations. Screen level metrics can be used to assess reproducibility at a screen level, being based on the spectrum of activity across assays. In this application of monitoring, screen level metrics are calculated giving the most weight to decisions on very low actives and very high actives. Decisions on very low actives are important due the high volume of experimental compounds being tested with no or very low biological activity, and thus, low business value. The decisions on very high actives are important due to the potential value in earnings for very active compounds. The screen level metrics are informative to managers.

7. Assay Level Decision Quality Metrics Calculations

The underlying philosophy of these metrics is simple. Assess the intrinsic activity of each standard item (i.e., unique assay/standard/rate combination) tested, based on multiple replicated runs. In this application of monitoring, the median percent control response is used as the measure of intrinsic or baseline activity. The baseline activity for each standard item is used to bin each item as active or inactive based on the assay promotion criteria. For example, a rate of a standard that gives an average response of 60% control would be binned as inactive, if the assay promotion criterion requires an average response greater than 80% control. Also, a subset of both the active and inactive bins with standard items whose baseline activity is either very low or very high can be created, thus identifying standard items that on average give either a very high or very low response. Initially, at least 12 runs of screening results for standard items are used to assess the baseline activity and 'bin' the standard items. Thereafter, periodically, (e.g., every year) this baseline assessment can be redone. However if a planned change in the screening process occurs, the baseline assessment must be recalibrated based on data from standard items that have be tested in the revised screening process.

A hypothetical example of setting the baseline for the assay level metrics is shown in Figure 6. The baseline activity for each standard item may be recalibrated periodically or when a bioassay procedure is changed.

An example of setting the baseline for an assay is shown in Figure 7. In this example, an assay tests four replicates of eight rates of five standards, a total of 160 test units, with each run. In automated early screening levels adding multiple standard items to the screening process may not create much additional work and can give much information about the reproducibility of the bioassay. Instead of attempting to assess the quality of an assay using the reproducibility of several replicates for each experimental compound, one can use the reproducibility of replicated standard items, tested run after run.

To calculate the metrics, on a run by run basis, the same promotion criteria is applied to each test unit of a standard item, just as if it were an experimental compound, assessing it as active or inactive. This assessment is compared to the designated intrinsic activity status, active or inactive, for that item. Correct and incorrect decisions for standard items in the active, inactive, very low active and very high active bins can be counted. Estimated error rates, based on these ‘surrogate’ experimental compounds are displayed as run charts, plotting either the percent correct decisions or percent incorrect decisions across time. Run charts are useful for identifying trends before they become problems.

Figure 8 displays the calculation of the estimated false positive (FP) rate for a run of the example assay shown in Figure 7. The FP estimate for a run is based on standard items whose baseline activity is less than the promotion criteria for that assay. For these ‘inactive’ standard items, the FP estimate for a run is the percent of ‘inactive standard items’ which gave an observed response over the promotion criteria. Figure 9 shows the calculation of the estimated false negative (FN) rate, based on response of standard items whose baseline activity is above the promotion criteria. In both Figure 8 and Figure 9, the numerators used in the calculations for FP and FN estimates are counts of observed data for that run that gave an unusual response for their ‘bin’ (i.e., FP or FN). The denominators are the count of standard items in each bin. Figure 10 displays run charts showing estimated percent correct decisions for both actives and inactives, based on assay level metrics. The number of standard items tested weekly in each category is also included in the legend. It is expected that observed responses for standard items with baseline activity near the promotion threshold will vary more than those standard items with very low activity or very high activity.

8. Use of Decision Quality Metrics

Decision Quality Metrics are useful to both the screen operators and management. Presentation of the Decision Quality Metrics in run charts provides a long-term process view allowing the early detection of trends and translating that variability into potential impact on error rates.

For example, if in one run many more compounds than usual are identified as active, a quick check on the Decision Quality Metrics for that screen will provide useful information about the accuracy of those identifications. If the Decision Quality Metrics show no shift for that run, one has more confidence that all the active assessments are correct. However, if that same run, there was an increase in the estimated false positive rate, with standard items that have very low activity showing increased activity, then perhaps some of those seemingly active compounds are really false positives. Further investigation into this shift in the sensitivity in the screen would be warranted in this case. While metrics monitor the health of assays, it is the quality management disciplines of procedural controls, change management and audits that continually improve the stability of repeated bioassays.

9. Summary

The Decision Quality Metrics translate observed variability of controls into potential impact on decision making. These metrics provide a more unbiased view of the health of a screening process than the percent active compounds, which is dependent on the selection of chemical compounds for testing. Based on the operating characteristic curves approach, these metrics work with the probability of a compound being passed as active, given the "intrinsic" level of activity of that compound. Various rates of commercial crop protection products are used as surrogates for experimental test compounds of varying activity levels. The information gained from the response of bioassays to these repeated stimuli is translated into estimated false positive and false negative rates. Run charts based on these metrics show decision reliability through time. This approach could be refined by updating the probabilities associated with each standard/rate combination using a moving or time series approach.

10. Acknowledgements

We would like to acknowledge Bruce Stanley, Jeffrey Wetherington and Steve Foor for their contributions in the development of the concepts and some of the early drafts of the figures presented in this paper. I would like to thank Bruce Stanley for his review of the paper.

References

Farnum, Nicholas R. *Modern Statistical Quality Control and Improvement.* California: Duxbury Press, 1994.

Zhang, J., Chung, T., Oldenburg, K. 1999. "A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays." Journal of Biomolecular Screening, Vol. 4(2), pp 67-73.

Figures and Tables

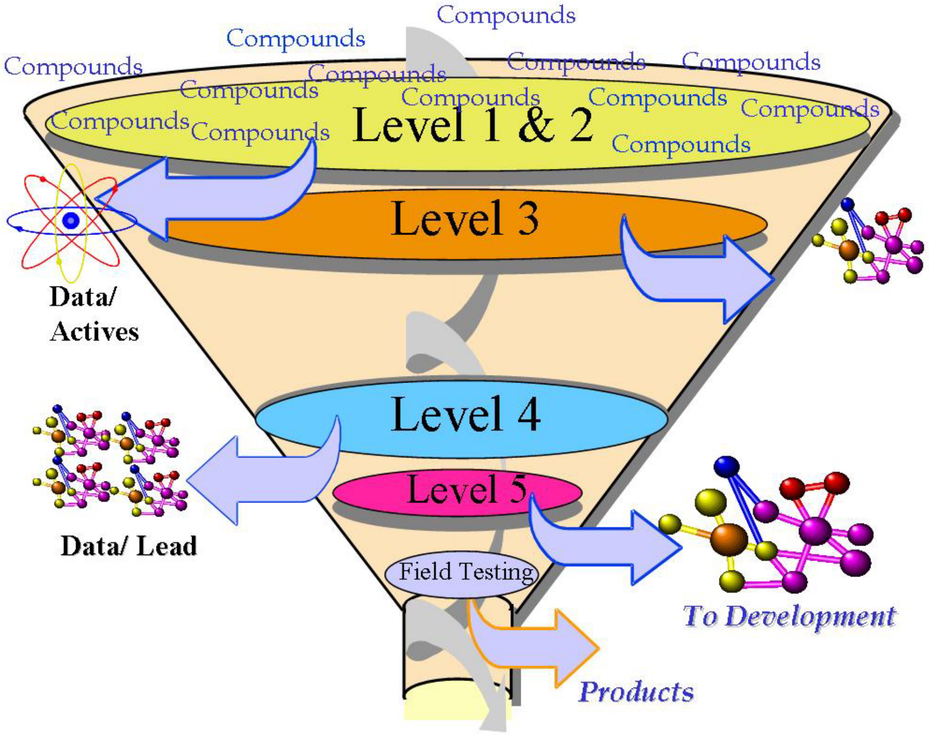


Figure 1. A graphical representation of the way experimental compounds may move through a tiered screening system. (Adapted from earlier version by J. Wetherington and S. Foor.)

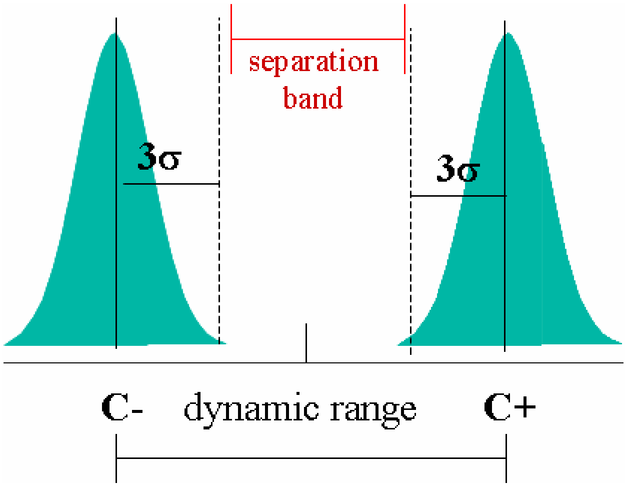


Figure 2. The Z-factor metric is based in the signal to noise ratio of the difference between the minimum and maximum signals.

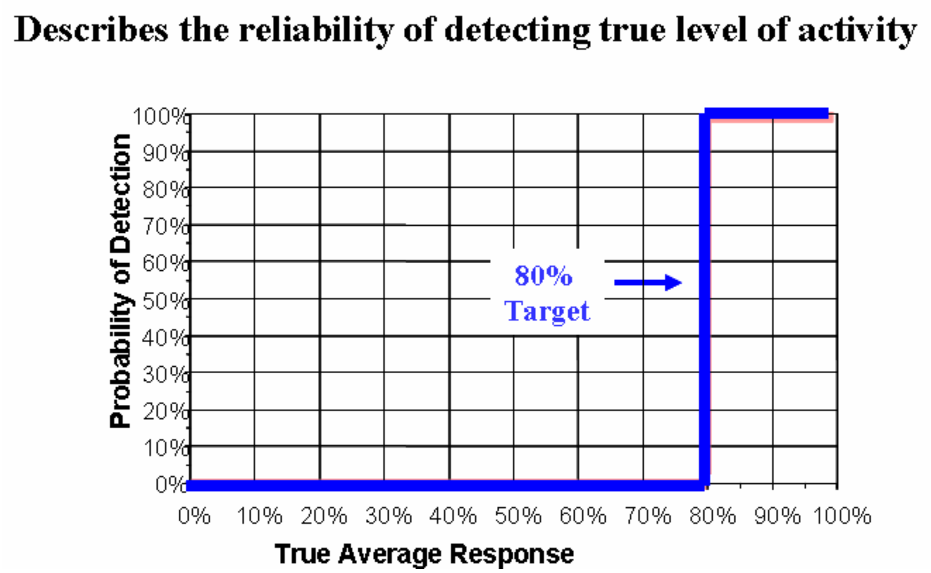


Figure 3. An ideal operating characteristic curve.

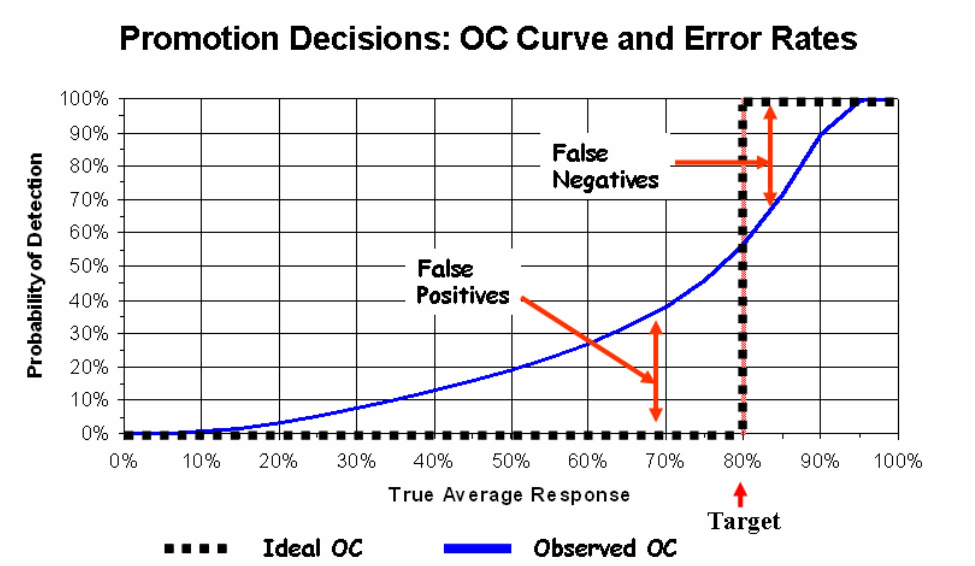
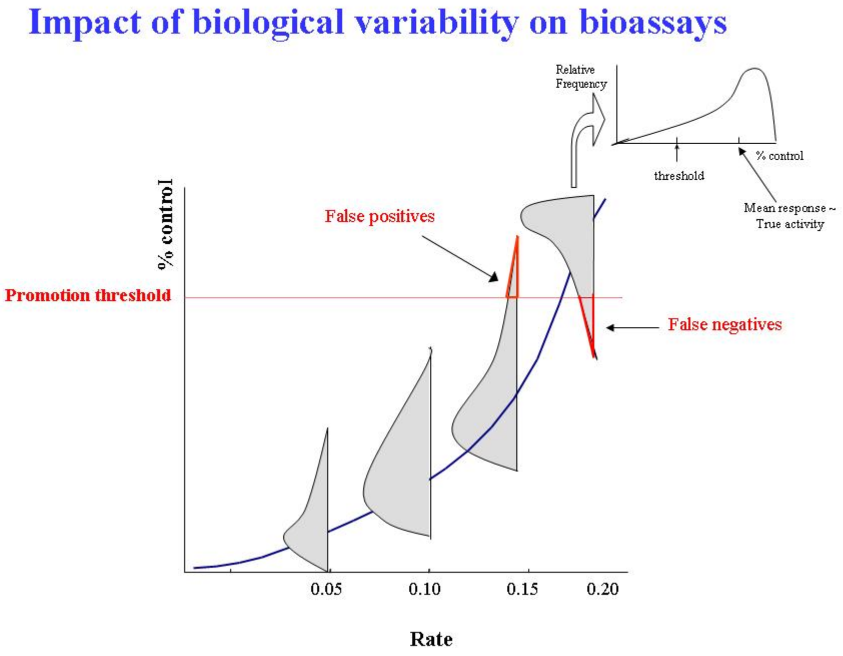
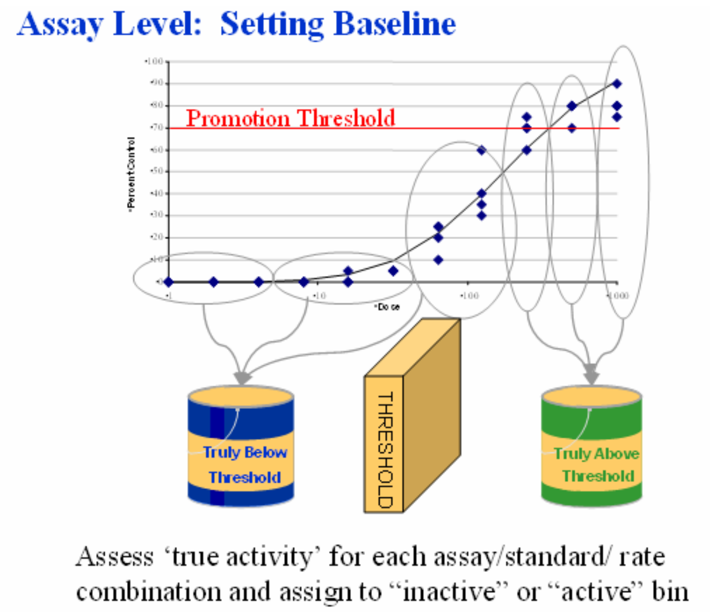


Figure 4. An operating characteristic curve, with 80% control as target activity. False positives are intrinsically inactive compounds identified by an assay as active. False negatives are intrinsically active compounds not identified as active by an assay. When a randomly selected compound has a uniform distribution of having any level of activity, the area under the operating characteristic curve, to the left of the x value of 80%, represents the false positive rate. The area above the operating characteristic curve, to the right of the x value of 80% represents the false negative rate.



Impact of biological variability on bioassays

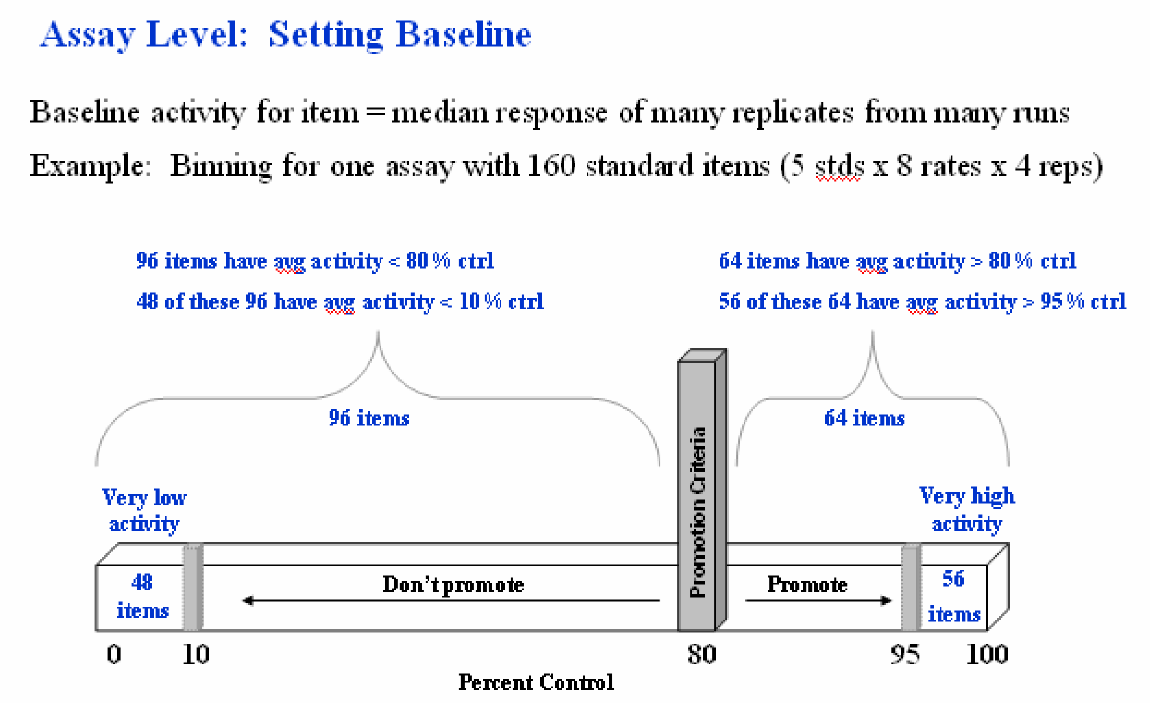
Figure 5. A hypothetical graphic demonstrating the impact of biological variability on error rates. Bioassays of compounds demonstrating intrinsic activity show a distribution of responses that follow a dose response curve. The dose response curve goes through the mean response for each rate tested. Replication, both within and across runs, at each rate tested would yield a distribution of responses. False positives are that part of a distribution of possible responses for a rate of a truly inactive compound that are observed with activity greater than the promotion criteria. False negatives are those realizations of a rate of a truly active compound that are under the promotion criteria.



Assay Level: Setting Baseline

Assess 'true activity' for each assay/standard/ rate combination and assign to "inactive" or "active" bin

Figure 6. A hypothetical example of setting the baseline for assay level metrics.



Assay Level: Setting Baseline

Baseline activity for item = median response of many replicates from many runs

Example: Binning for one assay with 160 standard items (5 stds x 8 rates x 4 reps)

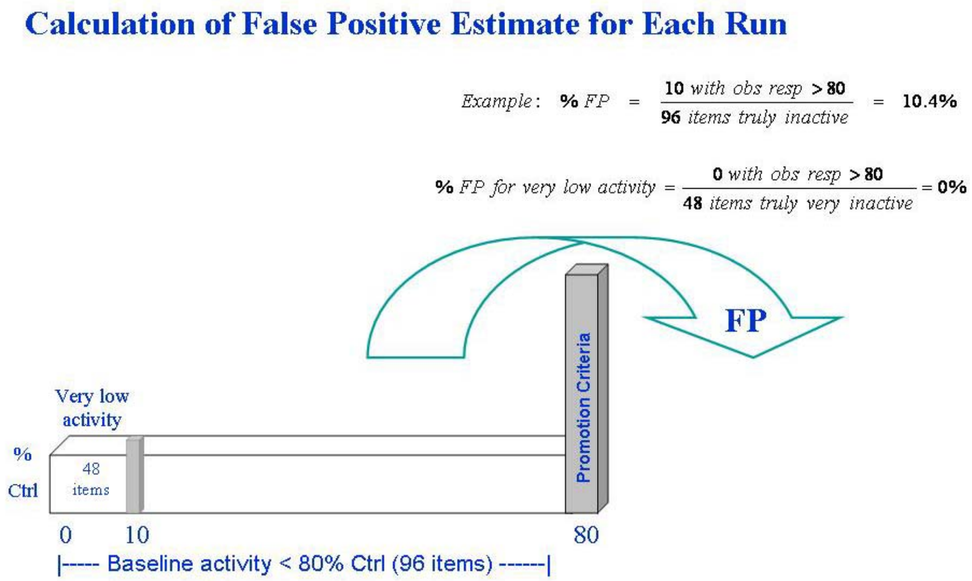
96 items have avg activity < 80 % ctrl

48 of these 96 have avg activity < 10 % ctrl

64 items have avg activity > 80 % ctrl

56 of these 64 have avg activity > 95 % ctrl

Figure 7. An example of setting the baseline for the assay level metrics.

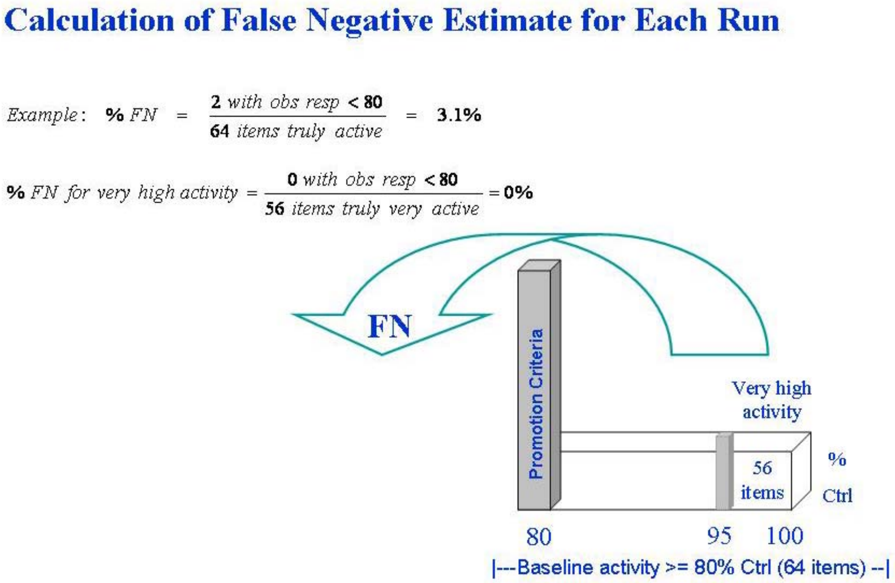


Calculation of False Positive Estimate for Each Run

$$\textit{Example}: \quad \%\, FP = \frac{\mathbf{10}\ \textit{with obs resp} > \mathbf{80}}{\mathbf{96}\ \textit{items truly inactive}} = \mathbf{10.4\%}$$

$$\%\, FP\ \textit{for very low activity} = \frac{\mathbf{0}\ \textit{with obs resp} > \mathbf{80}}{\mathbf{48}\ \textit{items truly very inactive}} = \mathbf{0\%}$$

Figure 8. False positive (FP) estimate is based on the response of standard items whose baseline activity is less than promotion criteria. To calculate FP estimate, for each run, calculate % standard items with baseline activity below promotion criteria which gave an observed response over the promotion criteria.

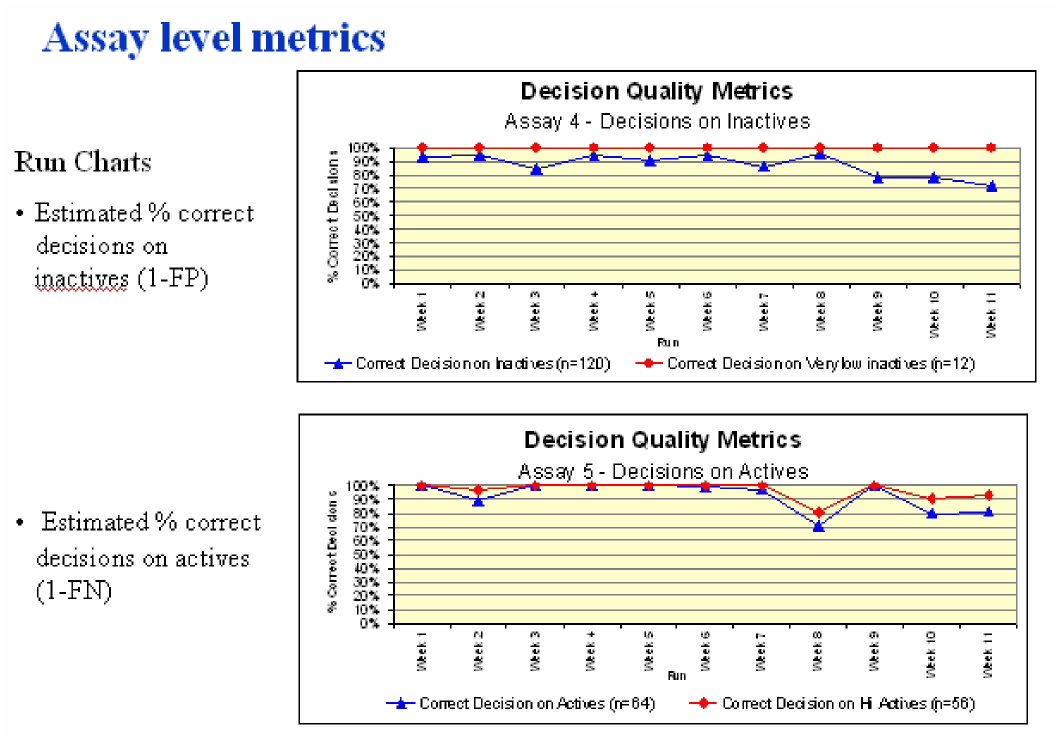


Calculation of False Negative Estimate for Each Run

$$\textit{Example}: \quad \%\, FN = \frac{\mathbf{2}\ \textit{with obs resp} < \mathbf{80}}{\mathbf{64}\ \textit{items truly active}} = \mathbf{3.1\%}$$

$$\%\, FN\ \textit{for very high activity} = \frac{\mathbf{0}\ \textit{with obs resp} < \mathbf{80}}{\mathbf{56}\ \textit{items truly very active}} = \mathbf{0\%}$$

Figure 9. False negative (FN) estimate is based on the response of standard items whose baseline activity is greater than promotion criteria. To calculate FN estimate, for each run, calculate % standard items with baseline activity above promotion criteria which gave an observed response under the promotion criteria.



Assay level metrics

Run Charts

- Estimated % correct decisions on inactives (1-FP)

- Estimated % correct decisions on actives (1-FN)

Figure 10. Example run charts showing estimated percent correct decisions on actives and inactives for the assay level metrics.