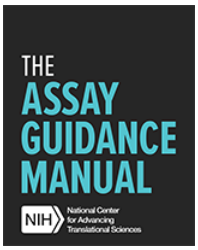
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Haas J, Manro J, Shannon H, et al. *In Vivo* Assay Guidelines. 2012 May 1 [Updated 2012 Oct 1]. In: Markossian S, Grossman A, Brimacombe K, et al., editors. Assay Guidance Manual [Internet]. Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences; 2004-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



In Vivo Assay Guidelines

Joseph Haas, MS,[✉1,*] Jason Manro, MS,[1] Harlan Shannon, PhD,[2] Wes Anderson, MS,[1,**] Joe Brozinick, PhD,[1] Arunava Chakravartty, PhD,[3] Mark Chambers, PhD,[1] Jian Du, PhD,[1] Brian Eastwood, PhD,[1] Joe Heuer, PhD,[1] Stephen Iturria, PhD,[4] Philip Iversen, PhD,[1,***] Dwayne Johnson, BS,[1] Kirk Johnson, PhD,[1] Michael O'Neill, PhD,[1] Hui-Rong Qian, PhD,[1] Dana Sindelar, PhD,[1] and Kjell Svensson, PhD[1]

Created: May 1, 2012; Updated: October 1, 2012.

Abstract

This document is intended to provide guidance for the design, development and statistical validation of *in vivo* assays residing in flow schemes of discovery projects. It provides statistical methodology for pre-study, cross-study (lab-to-lab transfers and protocol changes), and in-study (quality control monitoring) validation. Application of the enclosed methods will increase confidence in data from *in vivo* assays in the critical path and enable better decisions about SAR directions and compound prioritization. Screens using both single dose and multiple dose *in vivo* assays are discussed, along with acceptance criteria at different stages of validation.

Flow Chart

This document is intended to provide guidance for the development and statistical validation of *in vivo* assays residing in flow schemes of discovery projects. It provides statistical methodology for pre-study, cross-study (lab-to-lab transfers and protocol changes), and in-study (quality control monitoring) validation. Application of the enclosed methods will increase confidence in data from *in vivo* assays in the critical path and enable better decisions about SAR directions and compound prioritization.

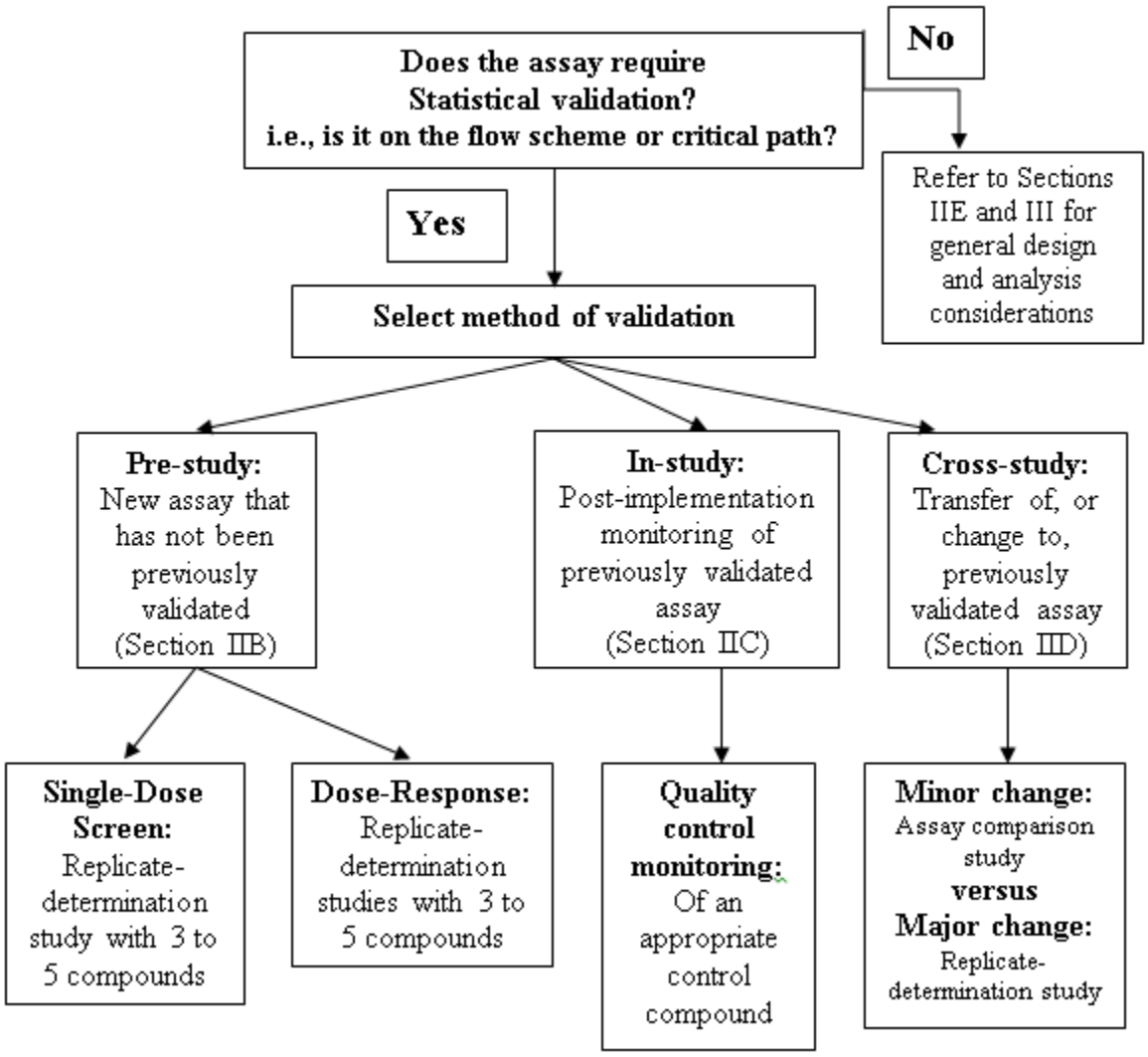
Author Affiliations: 1 Eli Lilly & Company, Indianapolis, IN. 2 Indiana University Purdue University at Indianapolis, Indianapolis, IN. 3 Abbott Laboratories, Chicago, IL. 4 Monsanto, Ankeny, IA.

✉ Corresponding author.

*Principal Author

**Co-Author

***Editor



The document is organized into three sections, and includes several examples. The Introduction (Section 1) provides general definitions of biological assays and provides general concepts in assay development and validation. Assay Validation (Section 2) contains the procedures and statistical details for *in vivo* assay validation, while Background Material (Section 3) covers general statistical concepts related to the design and analysis of *in vivo* experiments.

1. Introduction

This document is written to provide guidance to investigators who are developing and statistically validating *in vivo* assays for the evaluation of structure-activity relationships and/or compound collections to identify chemical probes that modulate the activity of biological targets. Specifically, this manual provides guidelines for:

a. Identifying potential assay formats of *in vivo* models compatible with **Single Dose Screens** (SDSs) and **Dose-Response Curves** (DRCs) for evaluating structure-activity relationships (SAR).
b. Statistical validation of the assay performance parameters (pre-study, in-study, and cross-study validation).
c. Optimizing assay protocols with respect to sensitivity, dynamic range, and stability.

1.1. General definition of biological assays

A biological assay is defined by a set of methods that produce a detectable signal allowing a biological process to be quantified. In general, the quality of an assay is defined by the *robustness and reproducibility* of this signal in the absence of any test compounds or in the presence of inactive compounds. This robustness will depend on the type of signal measured (biochemical, physiological, behavioral, etc.), and the analytical and automation instrumentation employed. The quality of the SDS is then defined by the behavior of this assay system when screened against a collection of compounds. These two general concepts, assay quality and screen quality, are discussed with specific examples in this manual.

1.2. General Concepts in Method (Assay) Development and Validation of an *In vivo* Model

The overall objective of any method validation procedure is to demonstrate that the method is *acceptable for its intended purpose*. Usually, the purpose is to determine the biological and or pharmacological activity of new chemical entities (NCE). The acceptability of a measurement procedure or bioassay method begins with its design and construction, which can significantly affect its performance and robustness.

The validation process originates during identification and/or design of a model and method development and continues throughout the assay life cycle (Figure 1). During method development, assay conditions and procedures are selected that minimize the impact of potential sources of invalidity (e.g. so-called false positives or false negatives) on the measurement of analyte or the biological end point (e.g. biochemical, physiological or behavioral changes). There are three fundamental general areas in method development and validation: (a) Pre-study (Identification and Design phase) validation (b) In-study (Development and Production phase) validation, and (c) Cross-validation or method transfer validation. These stages encompass the systematic scientific steps in an assay development and validation cycle.

1.2.1. Pre-study validation:

This validation occurs prior to implementing the assay. At this stage the choice of an assay format is made. Close attention must be paid to factors such as the selection of methods with appropriate specificity and stability. It is important that the assay be designed and the protocol written to facilitate the statistical analysis (i.e. appropriate design and analysis method, and adequate sample size), and that proper randomization techniques be used. This requires the generation and statistical analysis of confirmatory data from planned experiments to document that analytical results satisfy pre-defined acceptance criteria. Within-run statistical measures of assay performance such as the Minimum Significant Difference (MSD) for SDSs and Minimum Significant Ratio (MSR) for DRCs are calculated. If available, the assay sensitivity and pharmacology is evaluated using control compounds.

1.2.2. In-study validation:

These procedures are needed to verify that a method remains acceptable *during its routine use*. In order to compare data for compounds tested at different times, the pre-study statistical measures of assay performance (MSD or MSR) are updated to include between-run variability. Each run of the assay should contain appropriate {maximum and minimum} control groups or treatments to serve as quality controls of each run and to check overall performance. (Maximum and/or minimum quality controls differ conceptually from an “active” positive or negative controls; see Control Compounds or Treatment Groups.) This will allow the investigator to check for procedural errors and to evaluate stability of the method over time. Control Charts illustrates procedures which may be used to evaluate assay performance over time (i.e., control chart monitoring).

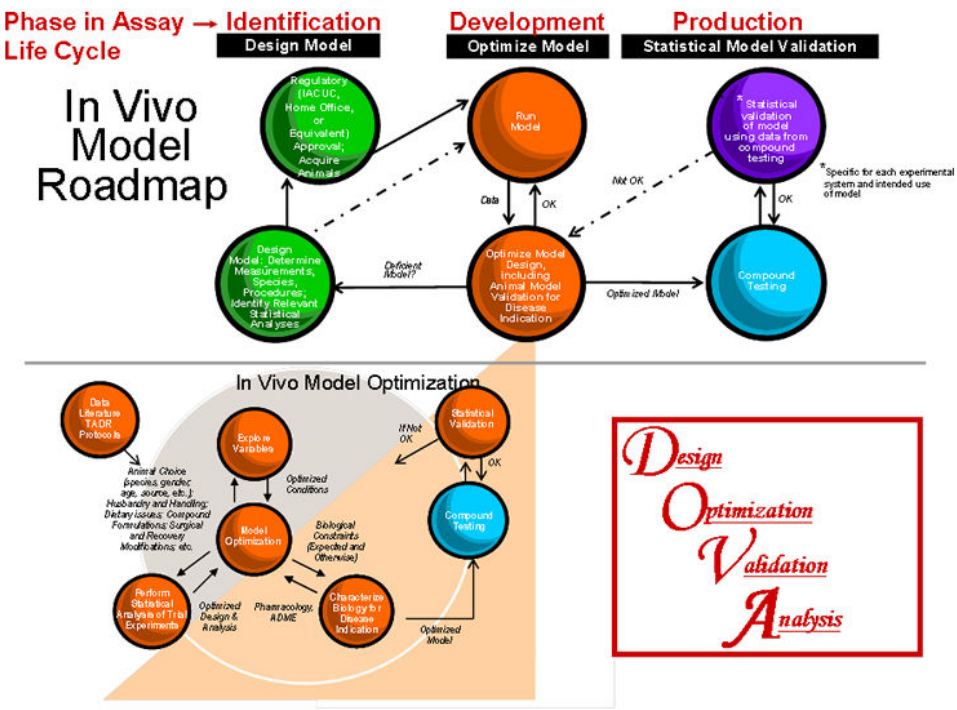


Figure 1: The *In vivo* Model Assay Development Cycle.

1.2.3. Cross validation:

This portion includes the *assay hand-off* from the individual investigator's team to another laboratory or a screening center. More broadly, this procedure is used at any stage to verify that an acceptable level of agreement exists in analytical results before and after procedural changes in a method as well as between results from two or more methods or laboratories. Typically, each laboratory assays a subset of compounds and the agreement in results is compared to predefined criteria that specify the allowable performance of the assay.

1.2.4. Resources:

The validation guidelines described here should be applicable to most *in vivo* models encountered in drug discovery research. However, situations could arise in which their verbatim application would be impractical given resource constraints, intended use of the assay, or other reasons. In these situations, and in general, the following principles should apply:

- Some form of statistical validation should always be performed and is better than no validation.
- The amount of resources, including time, spent on validation should be kept to a reasonably small fraction of the total resources to be used for testing compounds. What is "reasonable" will be have to be determined by the key personnel involved with each project.

These guidelines are intended to be used as "guidelines" and not exact "requirements." The specified assay performance measures serve to quantify how well the assay is performing and should be used to guide proper interpretation of the data. Determining whether an assay is "fit for its intended purpose" should be based on a combination of these assay performance measures and sound scientific judgment of the team.

2. Assay Validation Procedures

2.1. Overview

The statistical validation requirements for an assay vary, depending upon the prior history of the assay. The four main components of the statistical validation are:

1. Adequate study design and data analysis method.
2. Proper randomization of animals.
3. Appropriate statistical power and sample size.
4. Adequate reproducibility across assay runs.

Assays should be designed so that all biologically meaningful effects are statistically significant. In an exploratory study, this "meaningful effect" might correspond to any effect that is pharmacologically relevant. For a project/program team, it might correspond to an effect that meets the critical success factors (CSFs) defined in the compound development flow scheme. Power and sample size analyses are especially relevant for experiments that are designed to address key endpoints in a flow scheme. It is not acceptable to set a CSF equal to the effect size that is statistically significant since that effect size may or may not be biologically relevant. A CSF should be established based upon its biological relevance to the discovery effort. The assay is designed, optimized and validated so that biologically meaningful effects (e.g., CSFs) are statistically significant. Quantifying the reproducibility of an *in vivo* flow scheme assay will enable a team to discern whether compounds tested in different runs of the assay are exhibiting differential activity. This will result in better decisions about SAR directions and compound prioritization.

If the assay is new, or has never been previously validated relative to the targeted purpose, or mechanism of action, of the assay, then full validation should be performed. If the assay has been previously validated in a different laboratory, and is being transferred to a new laboratory, then a Replicate-Determination study (Pre-Study Validation) and a formal comparison between the new assay and the existing (old) assay (i.e., cross-study validation; Cross Validation) should be performed. If an assay is being transferred, it is considered validated if it has previously been assessed by all the methods in Section 2.2, and is being transferred to a new laboratory without undergoing any substantive changes to the protocol. If the intent is to store the data under the same Assay Method Version (or AMV) in an electronic database as the previous laboratory's AMV, then an assay comparison study (Cross Validation) should be done as part of the Replicate-Determination study. Otherwise only the intra-laboratory part of the Replicate-Determination study (Pre-Study Validation) is recommended.

If the assay is updated from a previous version run in the same facility then the requirements vary, depending upon the *extent of the change*. Major changes require a validation study equivalent to the validation of a new assay. Minor changes require bridging studies that demonstrate the equivalence of the assay before and after the change. See Section 2.4 - Cross Validation for examples of major and minor changes.

An assay methodology which has been previously validated for a different target or mechanism of action should be validated in full when used for a new or different target or mechanism as the variability, in particular, and reproducibility may be quite different for different mechanisms even though the methods may be very similar or identical. This concept is analogous to separately validating receptor binding assays for different receptors.

2.2. Pre-Study Validation: Replicate-Determination Study for Single-Dose Screens and Dose-Response Curves

2.2.1. Overview and Rationale

It is important to verify that assay results from multiple determinations or assay runs have acceptable reproducibility with no material systematic trends in the key endpoints. In this section, we define how to

quantify assay variability and determine assay equivalence. We also explain the rationale for the statistical methods employed in calculating reproducibility of activity and potency. *We strongly recommend consultation with a statistician before designing experiments to estimate variability, and the sources thereof, described below*. In particular, you should discuss with a statistician alternatives for assays with significant time, resource, or expenditure constraints as well as assays which will be used to test a minimal number of compounds to properly balance validation requirements with these constraints.

Replicate-Determination studies with two runs are used to formally evaluate the *within-run* assay variability (i.e., pre-study validation), or to formally compare a new assay to an existing (old) assay (i.e., cross-study validation). Replicate-Determination studies also allow a preliminary assessment of the *overall* or *between-run* assay variability, but two runs are not enough to adequately assess overall variability. In-study methods (In-Study Validation) are used to formally evaluate the overall variability in the assay in routine use. Note that the Replicate-Determination study is a diagnostic and decision tool used to establish that the assay is ready to go into production by showing that the endpoints of the assay are reproducible over a range of efficacies or potencies. It is not intended as a substitute for in-study monitoring or to provide an estimate of the *overall* MSD or MSR.

It may seem counter-intuitive to call the differences between two independent assay runs "within-run." However, the terminology results from the way those terms are defined. Experimental variation is categorized into two distinct components: *between-run* and *within-run* sources.

Consider the following examples:

- Between-run variation: If there is variation in the concentrations of the components in the vehicle between two runs then the assay results could be affected. However, assuming that the same vehicle is used with all compounds within the run, each compound will be equally affected and so the difference will only show up when comparing the results of two runs: one run will appear higher on average than the other run. This variation is called *between-run* variation.
- Within-run variation: If the concentration of one compound in the vehicle varies from the intended concentration (or dose) then all animals receiving that compound will be affected. However, animals receiving other compounds will be unaffected. This type of variation is called *within-run* as the source of variation affects different compounds in the same run differently. Therefore, it is necessary to compare results for one compound on each of two occasions.
- Some sources of variability affect *both within- and between-run variation*. For example, environmental conditions in an animal room have the potential to contribute to both types of variability. Suppose within a run of a particular *in vivo* assay, the temperature in the animal room exhibits spatial variation and animal response is sensitive to temperature. Animals will then respond differently depending upon their location in the room, and these differences are *within-run* as not all animals are equally affected. In comparison, suppose that during a particular run of the assay the room temperature on average is higher or lower than in previous runs. In this instance, the animals will respond differently on average during this run relative to other runs, and since all animals are affected this is *between-run* variation. Thus, a variable such as animal room temperature can be a source of both within- and between-run variation.

The *total variation* is the sum of both sources of variation. When comparing two compounds across runs, one must take into account both the within-run and between-run sources of variation. But when comparing two compounds in the same run, one must only take into account the within-run sources, since, by definition, the between-run sources affect both compounds equally.

In a Replicate-Determination study, the between-run sources of variation cause one determination to be on average higher than another determination. However, it would be very unlikely that the difference between the two determinations would be exactly the same for every compound in the study. These individual compound

"differences from the average difference" are caused by the within-run sources of variation. The higher the within-run variability the greater the individual compound variation within assay runs.

The analysis approach used in the Replicate-Determination study is to estimate and factor out between-run variability, and then estimate the magnitude of within-run variability.

Note: The between- and within-run sources of variability assessed during assay validation calculations apply to the treatment group means, not to the animal level data. Animal-to-animal variability is obviously present in *in vivo* experiments and this variability affects the reproducibility of the resulting treatment group means, but as shown subsequently, the animal-to-animal variability is not directly used in the calculation of the MSD and MSR. It is used to assess whether a data transformation is needed, and to assess sample size adequacy. See Section 2.5 for additional information about sources of variability in an *in vivo* assay.

2.2.2. Procedure (Steps)

All assays should have a reproducibility comparison (Steps 1 – 3). Single-dose screens should be validated separately from dose-response curve determinations, but it may be possible to validate both methods concurrently (consult with a statistician for options; also see Section 2.2.7 for an example). If the assay is to replace an existing assay using the same AMV code then an assay comparison study should also be done (Step 4).

1. Select a minimum of 3 to 5 compounds that have activities covering the effect range of interest and, if applicable, potencies that cover the dose-range to be tested. The compounds should be well spaced over these ranges.
2. All of the compounds should be tested in each of two runs of the assay.
3. Compare the two runs (as per Section 2.2.5 – 2.2.6.2.)
4. If the assay is to replace an existing assay:
    a. All compounds should be tested in a single run of the former assay as well as in two runs of the new assay. If a single run of the former assay already exists that meets the requirements in Step 1, it can be used for validation as long as the run is reasonably contemporaneous in time with the run of the new assay.
    b. Compare the results of the two assays, or labs, by analyzing the first run of the new assay with the single run of the former assay.

2.2.3. Summary of Acceptance Criteria (discussed in detail below)

1. For new assays, in Step 3 conduct reproducibility and equivalence tests for activity (single-dose screens and dose-response assays) and/or potency (dose-response assays) comparing the two runs. The assay should have an MSD < 20% and both Limits of Agreement on Differences (LsAd) between -20% and +20% for % activity. For potency results, recommendations are an MSR < 3 and both Limits of Agreement (LsA) between 0.33 and 3.0.
2. For assay transfer purposes, in Step 3 conduct reproducibility and equivalence tests for activity (single-dose screen and dose-response assays) and/or potency (dose-response assays) comparing the two runs in the new lab. The assay should have an MSD < 20% and both Limits of Agreement between -20% and +20% for % activity, and an MSR < 3 and both Limits of Agreement between 0.33 and 3.0 for potency.
3. For assay transfer purposes, in Step 4b conduct reproducibility and equivalence tests for activity and/or potency comparing the first run of the new lab to the single run of the old lab. The assays should have Limits of Agreement between -20% and +20% for % activity and Limits of Agreement between 0.33 and 3.0 for potency to be declared equivalent.

2.2.4. Notes

1. The Replicate-Determination study as laid out in Sections 2.2.2 and 2.2.3, assumes that the entire study is accomplished in just two runs of the assay (i.e. selected compounds are tested in each of two runs). While this design is preferred, it may not be possible for low-throughput assays in which only a small number of compounds can be tested per run. Alternatives are available in which each half of the replicate determination spans multiple runs (i.e., selected compounds are tested in each of a set of assay runs). Teams should discuss possible design alternatives and the appropriate analysis with their statistician. See also Section 2.2.7 for one possible alternative design and analysis.
2. *The acceptance criteria summarized above should be considered guidelines*, and may in fact be too stringent for some assays. Failure to meet the acceptance guidelines does not necessarily mean that the assay is unusable. Teams should consult with their statistician to understand the ramifications of missing the recommended criteria, and how it affects setting the CSF for the project and making decisions about compounds. For example, if the CSF is >80% and the MSD is 30%, then the assay will fail too many efficacious compounds, since even a 90%-active compound will fall below the CSF some of the time when the MSD is 30%. A more appropriate CSF in this situation might be 70 or even 60%. Furthermore, a 30% MSD indicates that a compound would need an activity test result of at least 80% to be considered to have differential activity from a compound with an activity test result of 50%. If the team desires increased power to discriminate among test compounds, the assay may need to be re-optimized to identify and reduce sources of variability (see Section 2.5 for additional details).
3. If a project is very new, there may not be 3 to 5 unique compounds with activity *in vivo* (where activity means some measurable activity above the minimum threshold of the assay). In that case it is acceptable to test compounds more than once, and/or at different doses, to get an acceptable sample size. For example, if there are only 2 active compounds then test each compound twice at the same two or three doses in each determination. However, when doing so, (a) it is important to biologically evaluate them as though they were different compounds, including independent sample preparation (i.e., weighing and solubilization), and (b) label the compounds and/or doses as "a", "b" etc. so that it is clear in the Replicate-Determination analysis which results are being compared across runs.
4. Dose-response assays need to be compared for both potency (ED_{50}) and efficacy (% maximum response). It is acceptable to use the same compounds for both single-dose and dose-response assays.

An assay may pass the reproducibility assessment (Steps 1-3 in the procedure [Section 2.2.2]), but may fail the assay comparison study (Step 4 in the procedure [Section 2.2.2]). The assay comparison study may fail either because of an absolute Mean Difference (MD) > 5% (or Mean Ratio (MR) different from 1), or a high MSD (or MSR) in the assay comparison study. If it's the former then there is an activity or potency shift between the assays. You should assess the values in the assays to ascertain their validity (e.g., which assay's results compare best to those reported in the literature?). If it fails because the assay comparison study MSD (or MSR) is too large (but the new assay passes the reproducibility study) then the old assay lacks reproducibility. In either case, if the problem is with the old assay, then the team should consider re-running key compounds in the new assay to provide comparable results to compounds subsequently run in the new assay.

2.2.5. Activity/Efficacy; Single Dose Screens

2.2.5.1. Analysis: Activity\Efficacy; Single Dose Screens

The points below describe and define the terms used in the acceptance criterion summarized in Section 2.2.3 and discussed in the Diagnostic Tests (Section 2.2.5.2). Individual animal data should be normalized to the positive (maximally effective) and negative (minimally effective) control averages to yield % activity for each animal. The computations that follow should then be made on the % activity averages for each compound, not the individual animal % activities.

1. Compute the difference in activity (first minus second) between the first and second determination for each compound. Let $\bar{d}$, and s_d be the sample mean and standard deviation of the difference in activity.
2. Compute the **Mean-Difference**: $MD = |\bar{d}|$ This is the average difference in activity between the two determinations.
3. Compute the **Difference Limits**: $DLs = \bar{d} \pm 2s_d/\sqrt{n}$, where n is the number of compounds. This is a 95% confidence interval for the Mean-Difference.
4. Compute the **Minimum Significant Difference**: $MSD = 2s_d$. This is the smallest activity difference between two compounds that is statistically significant.
5. Compute the **Limits of Agreement**: $LsAd = \bar{d} \pm 2s_d$. Most of the compound activity differences should fall within these limits (approximately 95%).
6. For each compound compute the **Difference** (first minus second) of the two activities, and the **Mean** activity (average of first and second).

Items 2-6 can be combined into one plot: the Difference-Mean plot (see example below).

2.2.5.2. Example

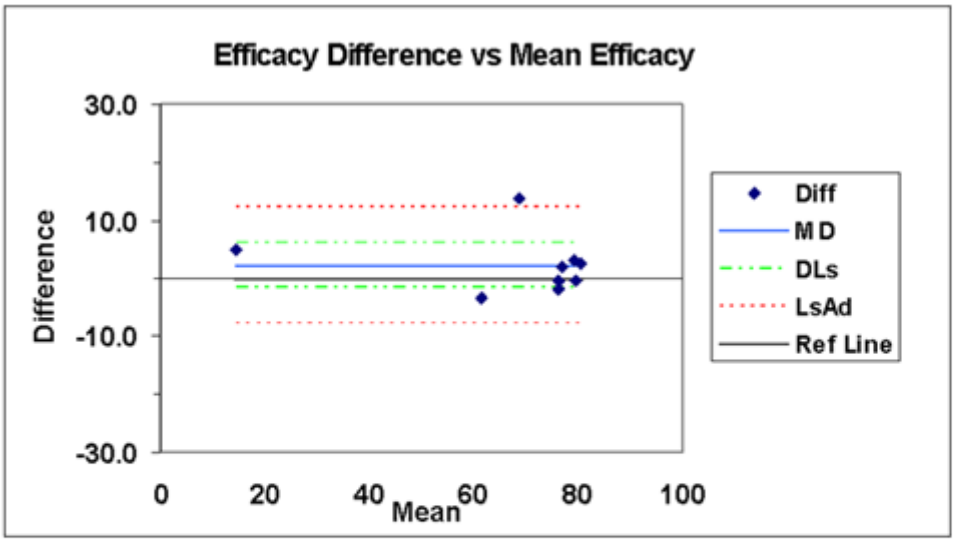
What follows is an example of an *in vivo* assay which measures the inhibition of EnzymeX Desaturase Index in rats, and which was transferred to Lab B from Lab A. Nine compounds were tested in each of two runs in Lab B with n=5 rats per compound. Table 1 displays the % inhibition averages of 5 rats per compound, along with the differences in compound averages. The MD, MSD, and Limits of Agreement are also shown which are computed from the average and standard deviation of the differences. This particular assay easily meets the acceptance criteria summarized in Section 2.2.3.

The compound averages can also be entered directly into the Replicate-Experiment template developed for in vitro assays and available on the NIH website. This template will perform the MD, MSD, and LsAd calculations and display them in the Difference-Mean plot.

Figure 2 shows the desired result of pure chance variation in the difference in activities between runs. The blue solid line shows the Mean Difference, i.e. the *average* relationship between the first and second run. The green long-dashed lines show the 95% confidence limits (or Difference Limits) of the Mean Difference. These limits should contain the value 0, as they do in this case. The red short-dashed lines indicate the Limits of Agreement between runs. They indicate the *individual compound* variation between the first and second run. You should see all, or almost all, the points fall within the red dashed lines. The lower line should be above -20, while the upper line should be below +20, which indicates a 20% difference between runs in either direction. The MD should be less than 5%, as it is in this example.

As this assay originated in Lab A, the lab comparison described in Step 4 of Section 2.2.2 was also completed. Eight of the nine compounds tested in the Lab B Replicate-Determination were also tested in Lab A. The eight compound averages of n=5 rats were entered into the template along with the corresponding eight compound averages from the first run in Lab B. Figure 3 shows the results, which meet the acceptance criteria summarized in Part 3 of Section 2.2.3.

Note: Although this assay achieved a successful transfer from Lab A to Lab B, demonstrating reproducibility within Lab B, as well as equivalence between labs, the validation has a shortcoming with respect to the selected compounds. *The selected compounds are not well spaced over the activity range of interest*, with most compound activities clustered together and falling well above the CSF of 50%. Since the reproducibility of a compound result from a given assay could depend on the activity level of the compound, it is very important to make an effort to cover the range of potential activity, and in particular to include compounds with activity bracketing the CSF (recognizing that this may be difficult depending on the maturity of a particular project).



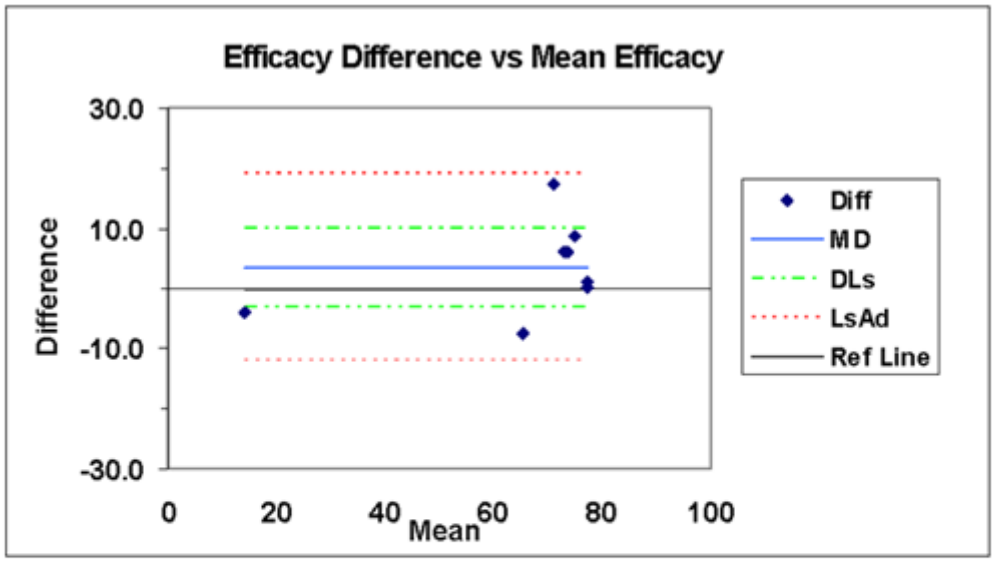
Equivalence Test Results

N	9
MD	2.20
DLs	-1.67
	6.07

Reproducibility Test Results

MSD (Within-Run)	10.06
LsAd	-7.87
	12.26

Figure 2: Result of reproducibility and equivalence tests for activity comparing the two runs in the new lab, Lab B. This example shows the desired result of pure chance variation in the difference in activities between runs.



Equivalence Test Results

N	8
MD	3.49
DLs	-3.03
	10.01

Reproducibility Test Results

MSD (Within-Run)	15.60
LsAd	-12.10
	19.09

Figure 3: Result of reproducibility and equivalence tests for activity comparing the first run of the new lab (Lab B) to the single run of the old lab (Lab A). This example shows results that meet the acceptance criteria.

Table 1: EnzymeX Desaturase Index Assay Replicate-Determination in Lab B.

Compound	Test1 %Inh	Test2 %Inh	Difference
A	76.19	76.68	-0.49
B	17.01	12.21	4.79
C	80.78	77.76	3.02
D	75.49	61.77	13.72
E	77.87	76.04	1.83
F	59.78	63.13	-3.35
G	79.47	79.84	-0.36
H	75.28	77.26	-1.98
I	81.92	79.30	2.61
MD = AVG of diffs		**2.20**	
MSD = 2 x STD of diffs		**10.06**	
LsAd = MD ± MSD (2.20 ± 10.06)		**-7.87**	**12.26**

2.2.5.3. Diagnostic Tests and Acceptance Criterion: Activity\Efficacy; Single Dose Screens

1. If the MSD ≥ 20% then there is poor agreement for **individual** compounds between the two determinations. This problem occurs when the *within-run* variability of the assay is too high. An assay meets the MSD acceptance criterion if the (within-run) MSD < 20%.
2. If the Difference Limits do not contain the value 0, then there is a statistically significant **average** difference between the two determinations. Within a lab (Step 3) this is due to high *between-run* assay variability. Between labs (Step 4), this could be due to a systematic difference between labs, or high between-run variability in one or both labs. Note that it is possible with a very "tight" assay (i.e., one with a very low MSD) or with a large set of compounds to have a statistically significant result for this test that is not very material, i.e., the actual MD is small enough to be ignorable. If the result is statistically significant then examine the MD. If it is between -5% and +5% then the average difference between runs is deemed immaterial.
3. The MD and the MSD are combined into a single interval referred to as the **Limits of Agreement**. An assay that either has a high MSD and/or an MD different from 0 will tend to have poor agreement of results between the two determinations. An assay meets the Limits of Agreement acceptance criterion if both the upper and lower limits of agreement are between -20 and +20.

2.2.6 Potency; Dose-response curves

2.2.6.1. Analysis: Potency; Dose-response curves

The points below describe and define the terms used in the acceptance criterion summarized in Section 2.2.3 and discussed in the Diagnostic Tests (Section 2.2.6.2). See Section 3.5 for guidelines for fitting *in vivo* dose-response curves.

1. Compute the difference in log-potency (first minus second) between the first and second determination for each compound. Let $\bar{d}$, and s_d be the sample mean and standard deviation of the difference in log-potency. Since ratios of ED_{50} values (relative potencies) are more meaningful than differences in potency (1 and 3, 10 and 30, 100 and 300 have the same ratio but not the same difference), we take logs in order to analyze ratios as differences.
2. Compute the **Mean-Ratio**: $MR = 10^{\bar{d}}$. This is the geometric average fold-difference in potency between two determinations.
3. Compute the **Ratio Limits**: $RLs = 10^{\bar{d} \pm 2s_d/\sqrt{n}}$, where n is the number of compounds. This is the 95% confidence interval for the Mean-Ratio.
4. Compute the **Minimum Significant Ratio**: $MSR = 10^{2s_d}$. This is the smallest potency ratio between two compounds that is statistically significant.
5. Compute the **Limits of Agreement**: $LsA = 10^{\bar{d} \pm 2s_d}$. Most of the compound potency ratios (approximately 95%) should fall within these limits.
6. For each compound compute the **Ratio** (first divided by second) of the two potencies, and the **Geometric Mean** potency: $GM = \sqrt{\text{first} \times \text{second}}$.

Items 2-6 can be combined into one plot: the Ratio-GM plot. The plot is very similar to the Difference-Mean plot described previously in Section 2.2.5 except that both axes are on the log scale instead of the linear scale.

2.2.6.2. Diagnostic Tests and Acceptance Criterion: Potency; Dose-response curves

1. If the MSR ≥ 3 then there is poor agreement for **individual** compounds between the two runs. This problem occurs when the *within-run* variability of the assay is too high. An assay meets the MSR acceptance criterion if the (within-run) **MSR < 3.**
2. If Ratio Limits do not contain the value 1, then there is a statistically significant **average** difference between the two determinations. Within a lab (Step 3) this is due to high *between-run* assay variability.

Between labs (Step 4), this could be due to a systematic difference between labs, or high between-run variability in one or both labs. Note that it is possible with a very "tight" assay (i.e., one with a very low MSR), or with a large set of compounds, to have a statistically significant result for this test that is not very material, i.e., the actual MR is small enough to be ignorable. If the result is statistically significant then examine the MR. If it is between 0.67 and 1.5 then the average difference between runs is less than 50% and is deemed immaterial. Note that there is no direct requirement for the MR, but values that are outside 0.67 and 1.5 are unlikely to pass the Limits of Agreement criterion in step 3 below.

3. The MR and the MSR are combined into a single interval referred to as the **Limits of Agreement**. An assay that either has a high MSR and/or an MR different from 1 will tend to have poor agreement of results between the two determinations. An assay meets the Limits of Agreement acceptance criterion if both the upper and lower limits of agreement are between 0.33 and 3.0.

2.2.7. Other Approaches for Assessing Reproducibility

2.2.7.1 Retrospective Assessment of Reproducibility

The preferred method for validating an in-vivo assay is by the Replicate-Determination Study described in Sections 2.2.1 – 2.2.6.2. This prospective approach provides an up-front assessment of the capability of an assay to reproducibly identify active compounds prior to placing the assay into production.

It is also possible to validate an assay retrospectively, that is, to collect validation data while screening test compounds. For example, a Single Dose Screen that has been adequately designed and powered, and that will use proper randomization and analysis techniques could be placed immediately into production. One could then include a "quality control" (see Sections 2.3.1 and 3.1 for a description) in each run of test compounds in order to assess the reproducibility component of assay validation. After six or more runs of the assay, the quality control data could be used to calculate an MSD for the assay.

One obvious risk in using the retrospective approach is that six or more runs of test compound data will be generated before knowing whether or not the reproducibility of the assay is adequate for a team's needs. This can be an efficient approach provided the resulting MSD is acceptably small; if it is not, the assay may need to be re-optimized to identify and reduce sources of variability (see Section 2.5 for additional details).

We strongly recommend consultation with a statistician before embarking on this approach to assessing the reproducibility component of assay validation. Teams should also consult a statistician regarding the calculation and interpretation of the MSD once the data are collected. Note also that this approach produces an "overall" MSD (see Section 2.1 for a description), since in addition to within-run variation, the resulting MSD also encompasses between-run variation from more than six runs.

2.2.7.1.1. Example

What follows is an example of an *in vivo* assay run in Lab B, which measures the inhibition of EnzymeX Desaturase Index in rats, and which was transferred to Lab B from Lab A and discussed in Section 2.2.5. The assay was retrospectively validated in Lab A by including a quality control in each of seven runs. These data are shown in Figure 4, which displays % inhibition data for individual animals as well as the averages for each run (connected by the blue line in the plot). The overall MSD of 16% is calculated from the standard deviation of the seven % inhibition averages and meets the validation criteria of MSD < 20%. The variability chart also indicates a decrease in animal-to-animal variation over time, as well as a slight increase in the activity of the quality control over time.

The % inhibition averages can also be entered into the control chart template (see Section 2.3.2 for additional details), which will calculate the overall MSD and allow monitoring of the averages over time. Output from the template is displayed in Figure 5.

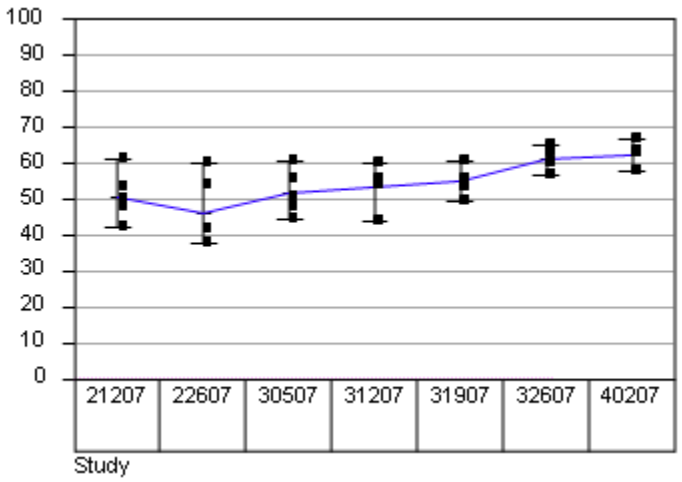


Figure 4: Variability chart of EnzymeX Desaturase Index versus Study.

2.2.7.2. Concurrent Validation of SDSs and DRCs

As mentioned in Section 2.2.2, SDSs should be validated separately from DRCs, but it is possible to validate both methods concurrently. Rather than beginning with the SDS validation via a Replicate-Determination Study (described in Section 2) using several compounds at a single dose, one could proceed directly to the DRC Replicate-Determination Study using a smaller number of compounds tested in dose-response format. The pairs of % activity determinations for the individual doses of each compound can be used to calculate an MSD that is applicable to the assay in single-dose format, while the pairs of ED_{50} determinations can be used to calculate an MSR that is applicable to the assay in dose-response format.

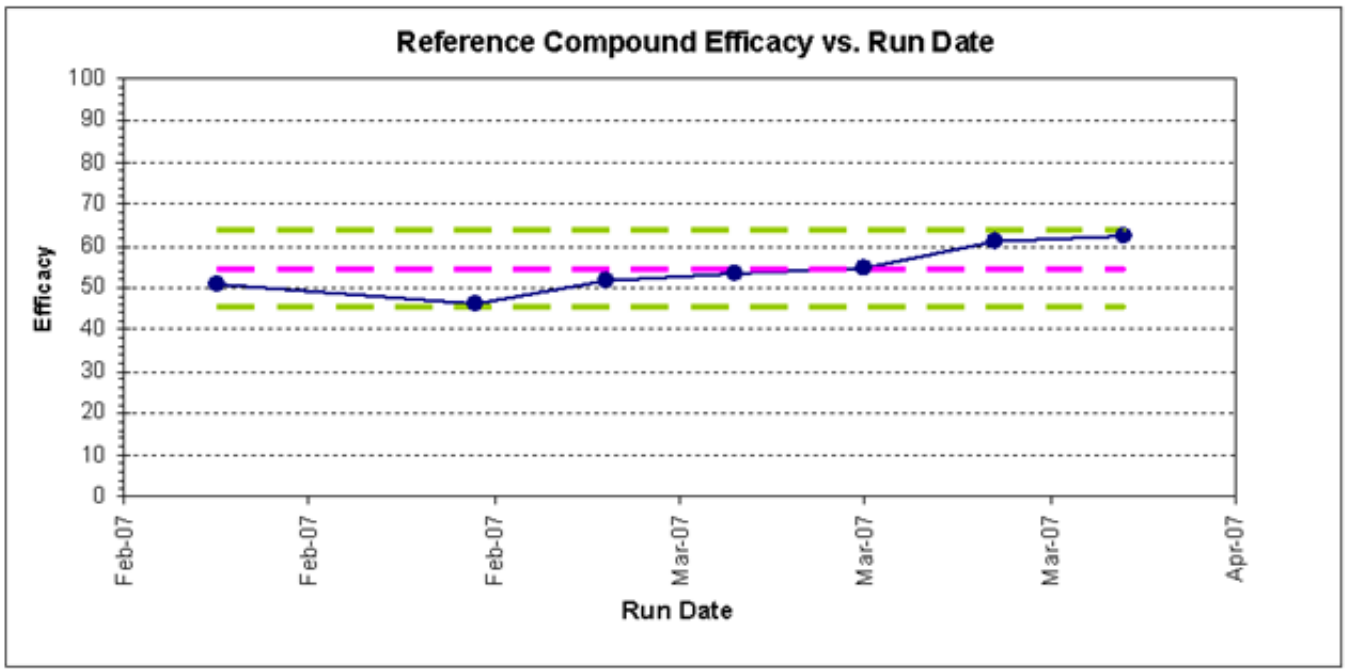
2.2.7.2.1. Example

The Capsaicin-Induced Eye Wipe Assay evaluates the ability of compounds to alleviate conjunctival irritation caused by exposure to capsaicin, as measured by inhibition of eye wiping in rats. The assay was developed and validated in Lab A, and upon completion of these activities, the assay was transferred to Lab B. Since the operator in the receiving laboratory was well-trained with respect to running the assay, and since no other major changes to the protocol were made, a bridging study or "Single-Determination Study" was utilized to validate the transfer (see Section 2.4 for additional details).

Three compounds were selected and tested in dose-response format with n=6 rats per each of four doses plus a capsaicin control for a total of 30 animals. The three compounds were tested once in the originating lab and once in the receiving lab. Percent inhibitions of the capsaicin-induced numbers of eye wipes are displayed in the variability chart (Figure 6). One compound was lost for technical reasons in one of the labs, so only two are charted.

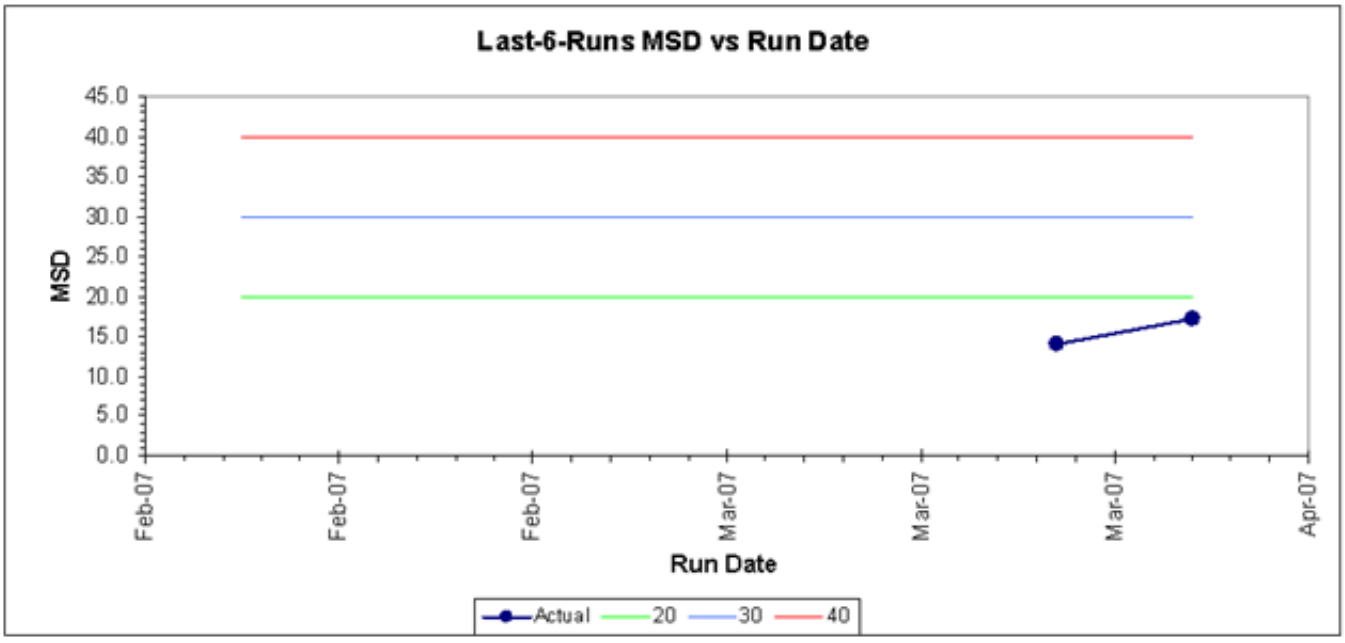
The variability chart indicates similar-dose response relationships for the two compounds between labs. There does however appear to be a consistent shift in the % inhibition for the individual compounds between labs. Also note that there is not complete overlap in the tested doses between labs, but each lab has four doses plus capsaicin available for estimation of the ED_{50} and 6 pairs of single-dose % inhibitions for calculation of the Mean Difference, Minimum Significant Difference, and Limits of Agreement.

This Single-Determination Study is also an example in which each lab's set of three determinations spanned multiple runs of the assay (i.e., each compound was tested in a separate run) as described in Section 2.2.4 Note 1. Teams should consult their statistician for appropriate MD, MSD, and LsAd calculations when determinations span multiple runs.



Control Limits

Mean (Center Line)	54.38
Upper Control Limit	63.60
Lower Control Limit	45.16



Mean Summary		**MSD Summary**	
High	62.48	High	17.15
Low	46.20	Low	14.02
Overall	54.38	Overall	16.31
Current	62.48	Current	17.15

Figure 5: Output from entering percent inhibition averages into the control chart template.

For the data in this example, the MD is 17%, the MSD is 11%, and the LsAd are 5.1% and 28%. While the MSD is excellent at 11%, the LsAd do not fall between -20% to +20 due to the 17% average shift in % inhibition between labs. While ED_{50}s from just two compounds are not enough to calculate an MSR, the ED_{50}s estimated from a four-parameter logistic fit also exhibited a shift in potency between labs, but the magnitude of the shift was less than 3-fold for both compounds.

Even though the LsAd did not meet the within ±20% criterion, the transfer validation was accepted and the assay was implemented in the receiving laboratory. This decision was based on discussion with scientists regarding possible explanations for the shift, as well as implications for the projects the assay supports. While the LsAd failed the ±20% criterion, the limits did fall within ±30%, and the ED_{50}s from the two compounds fell within 3-fold, which taken together was considered acceptable for the teams' needs. No MSR could be calculated, but since each compound's dose-response curve consisted of 30 independently tested animals (thereby providing confidence in the resulting ED_{50}s), and the ED_{50}s agreed to within 3-fold, the transfer was considered to be validated in both single-point and dose-response formats. Close monitoring of a quality control was also recommended for the assay going forward.

Another alternative approach is to include the dose used for the single-dose screen in dose-response determinations, and using the resulting repeats to calculate an MSD.

2.3. In-study Validation (Single-Dose Screens)

As mentioned previously in Section 2.2.1, the Replicate-Determination study is used to formally evaluate *within-run* assay variability. The study also provides a preliminary assessment of the *overall* assay variability, but typically does not involve enough assay runs to adequately assess *overall* variability. Overall assay variability consists of both within-run and *between-run* sources of variability, and hence needs to be estimated from several runs of the assay.

Post-implementation monitoring of the measured activity of a control compound that is regularly included in runs of the assay is an effective means for estimating the overall variability in the assay, which can then be used to calculate an overall MSD. The resulting MSD will likely be larger than the Replicate-Determination MSD, since the between-run variability is now included, but it is the appropriate MSD for comparing and prioritizing compounds tested in different runs of the assay.

Continuous monitoring of a regularly tested control compound is also an effective means of tracking assay performance over time, thereby ensuring high quality data for compound prioritization. Control charting the activity of the same compound over time ensures that the activity remains stable without appreciable "assay drift," and enables identification of suspect runs. Tracking assay performance over time will also ensure that the reproducibility of the assay (i.e., the MSD) remains at an acceptable level.

2.3.1. Control Compounds or Treatment Groups

Control compounds in *in vivo* experiments serve three purposes: (i) as comparator for the test compounds, (ii) to normalize responses across assay runs, and (iii) as a quality control marker. A ***negative* (minimally effect) *control*** is typically included in all runs of an assay and serves as the comparator for test compounds. A ***positive* (maximally effective) *control*** is also often included to establish that the assay is working and to normalize the response over assay runs. While these so-called minimum and maximum controls can be monitored over time to ensure that adequate separation is maintained, they may not be the best choice for monitoring assay performance and reproducibility as described in Section 2.3. Since these controls lie at the extremes of the dynamic range of the assay, they may not accurately represent the level of variability present within the interior of the range (i.e., assay variability may be different for low, medium, and high activity levels). Since interest often lies in prioritizing compounds whose activities fall in the middle to upper end of the dynamic range (but not necessarily at the top of the range), an additional control compound with a normalized (to positive and negative

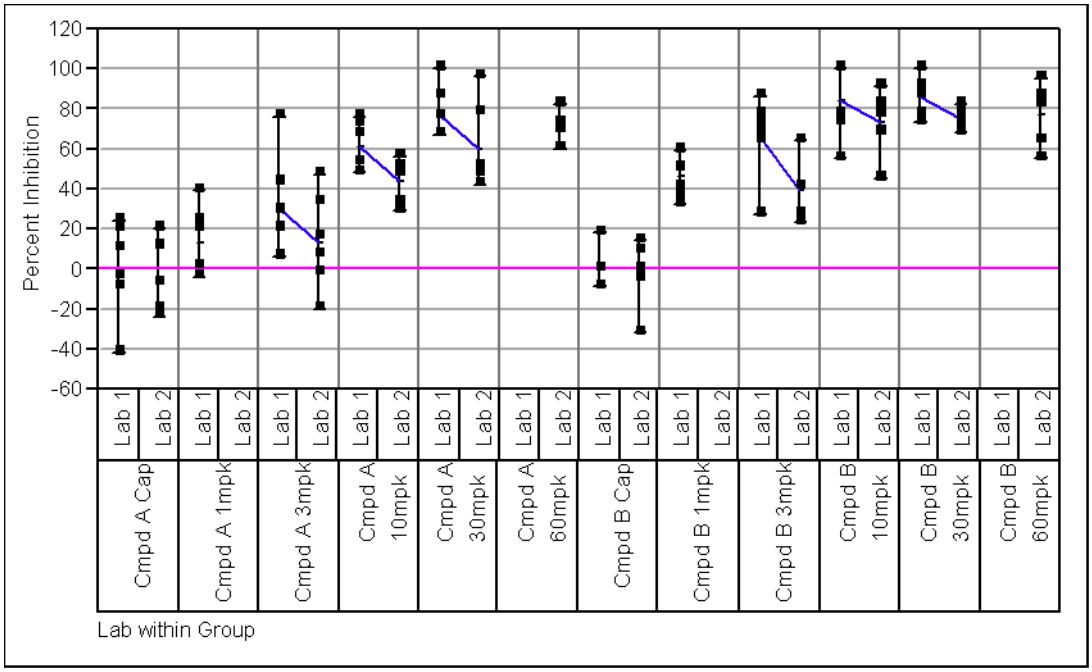


Figure 6: Variability chart for percent inhibition in a Single-Determination Study.

controls) activity of 50 to 70% should also be included to adequately monitor assay performance and reproducibility. This ***quality control*** should be included in each run of the assay, if possible, but could be included periodically (e.g. every other or every third run) if practical limitations (cost, time, resources etc.) exist for the assay.

2.3.2. Control Charts

Monitoring assay performance can be accomplished with a simple scatter plot of the % activity of the quality control versus run date, which should be updated after each run of the assay. Assay drift (trends up or down) can be identified visually, and problems investigated and corrected as they occur. Suspect runs should also be investigated and repeated if warranted.

Full-fledged control charts that calculate and display control limits are preferred, since the limits can assist with identifying trends or steps in the data. Control chart monitoring templates that produce scatter plots, as well as the preferred control charts, for *in vivo* assays are available (check with your statistician). Both templates can be used for *in vivo* data by first normalizing the individual animal data to the positive and negative control averages in each run, and entering the % activity averages for the quality control from each run into the templates.

After six runs, both templates will also calculate the overall MSD, which incorporates both within-run and between-run sources of variability (in contrast to the Replicate-Determination MSD, which encompasses only within-run variability). As such, it is expected that inclusion of the within-run variability will inflate the overall MSD as compared to the Replicate-Determination MSD. Teams should consult with their statistician about the interpretation of the overall MSD and how it affects comparing activities of compounds (and subsequent prioritization) tested in different runs of the assay. After each subsequent run, a running MSD (i.e., last-6-runs MSD computed from last six runs) is also computed and displayed graphically to enable monitoring of the reproducibility of the assay over time to ensure that it remains at an acceptable level.

2.3.2.1. Example

What follows is an example of an *in vivo* assay which measures the inhibition of EnzymeX Desaturase Index in DIO rats, and which was discussed previously in Section 2.2.5. A control compound was included in 14 runs of the assay. After normalizing the individual animal EnzymeX DIs to the positive and negative controls within each run, the average % inhibition (n=5 animals per group) of EnzymeX DI for the control compound in each run was entered into the control chart template. The output is shown in Figure 7.

Note the very stable activity of the quality control over the first 14 runs of the assay in the control chart of % activity versus run date. The % activity of the quality control tracks closely to the average of 79% over the 14 runs, with no material drifts or jumps in activity apparent in the chart.

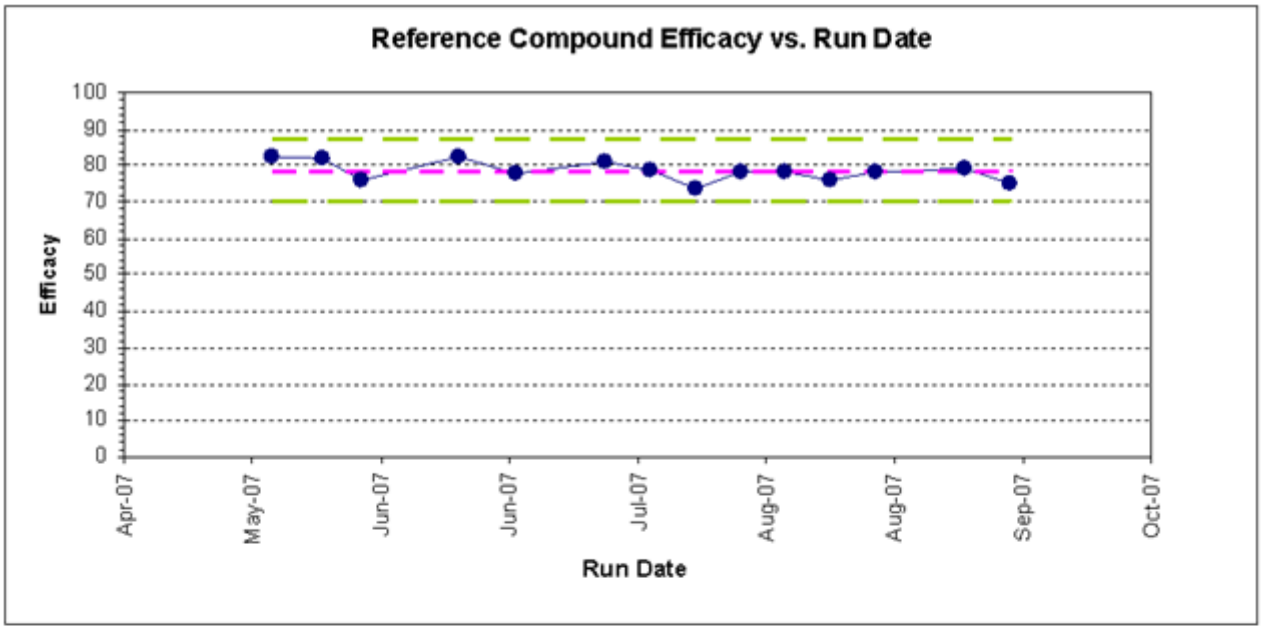
Also note the very stable level of reproducibility in the Last-6-Runs MSD versus run date chart (in fact, the reproducibility looks to be improving slightly over time). The running MSDs are well below 20% (actually fall below 10%), and the overall MSD is excellent at 7.6%, suggesting plenty of power to discriminate among compounds tested in different runs of the assay. However, it is important to note that the overall MSD of 7.64, which incorporates both within-run and between-run variability, is actually smaller than the Replicate-Determination MSD of 10.06 (Section 2.2.5), which incorporates only the within-run variability. This could be the result of the operator simply getting better at running the assay (the slight improvement in the running MSDs noted above provides some evidence for this), or it could be due to the fact that the quality control is quite active. While not maximally effective, the % activity of the quality control was certainly outside the recommended range of 50 to 70% activity for a quality control. Since variability (and hence reproducibility) may depend on the activity level of the control (with more active compounds expected to be less variable), the 79% active control may in fact be overestimating the true reproducibility of the assay.

2.4. Cross Validation: Bridging Studies for Assay Upgrades/Minor Changes versus Major Changes

Sections 2.2 and 2.3 cover the validation of entirely new assays, or assays that are intended to replace existing assays. The replacement assays are "different" from the original assay, either because of facility changes, personnel differences, or substantively different measurement and recording or automation equipment. Assay upgrades and changes occur as a natural part of the assay life cycle. Requiring a full validation for every conceivable change is impractical and would serve as a barrier to implementing assay improvements. Hence, full validation following every assay change is not recommended. Instead, bridging studies or "mini-validation" studies are recommended to document that the change does not degrade the quality of the data generated by the new assay. In addition, if the assay is to report results in an Assay Method Version (AMV) previously reported into by another assay then it has to be verified that the two labs produce equivalent results.

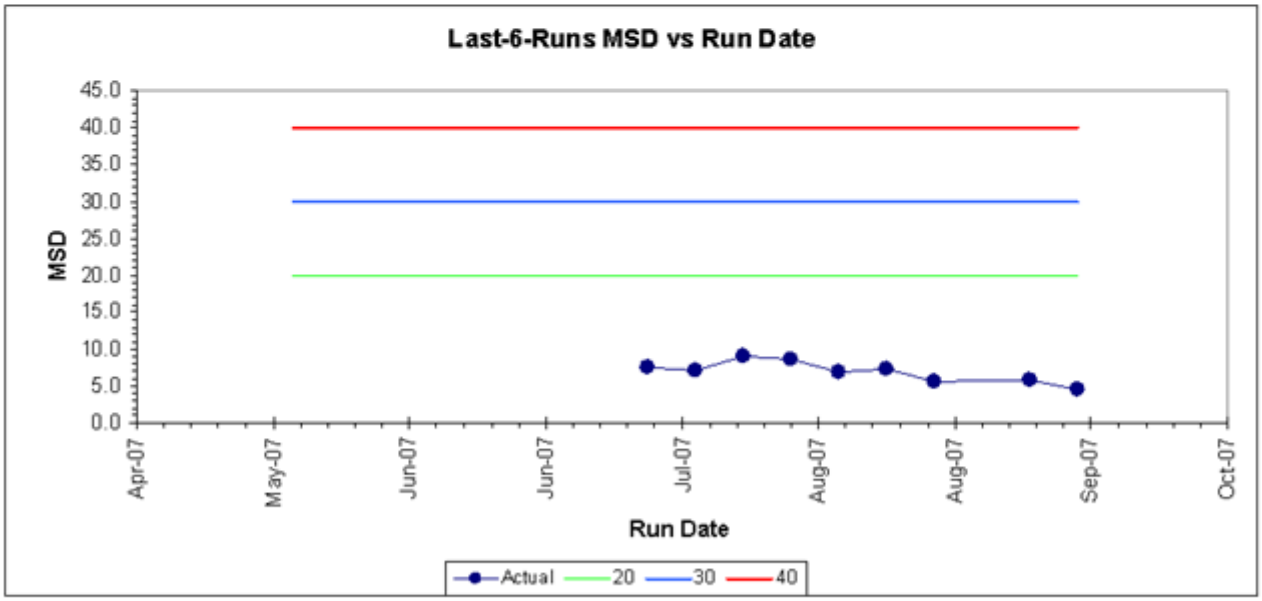
The level of validation recommended has two tiers, either a Single-Determination Study (Tier I), or a Replicate-Determination Study (Tier II) similar to the full validation package of Section 2.2. Examples of changes within each Tier are given below, along with the recommended validation study for that tier. Note that if the study indicates the change will have an adverse impact on assay quality (i.e., the study indicates there are problems), then the cause should be investigated, corrected, and a full validation should be done. If the results from that study indicate the assays are not equivalent, but the new assay is acceptable, then a new AMV should be established for the assay.

The following guidelines apply principally to changes in biological components of the protocol. If changes are made to the data analysis protocol then these can ordinarily be validated without generating any new data, but rather by comparing the results using the original and new data analysis protocols on a set of existing data. Discuss any changes with a statistician. If changes are made to both the data analysis and biological components



Control Limits

Mean (Center Line)	78.54
Upper Control Limit	87.24
Lower Control Limit	69.83



Mean Summary		**MSD Summary**	
High	82.46	High	9.07
Low	73.77	Low	4.45
Overall	78.54	Overall	7.64
Current	75.31	Current	4.45

Figure 7: Output from entering average percent inhibition of EnzymeX DI into the control chart template.

of the protocol then the appropriate tier should be selected according to the severity of the biological change as discussed below. The data analysis changes should be validated on the new validation data.

2.4.1. Tier I: Single Step and/or Minor Changes to the Assay

Tier I modifications are single changes in an assay such as a change in the person running the assay, an assay condition, instrumentation, or to a *reagent,* that is made either to improve the assay quality or increase the capacity without changing the assay quality. The changes can also be made for reasons unrelated to assay throughput or performance (e.g., change of a supplier for cost savings). Examples of such changes are

- Changes in supplier of animals
- Change in barrier facility used by the supplier
- Changes in recording instruments with similar or comparable electronics. E.g.: blood pressure recording instruments, clinical chemistry equipment, HPLCs, spectrophotometers, behavioral testing equipment such as locomotor activity instruments or operant conditioning chambers. A performance check for signal dynamic range, and signal stability is recommended prior to switching instruments.
- For *ex vivo* analysis of tissues, changes in liquid handling equipment with similar or comparable volume dispensing capabilities. Volume calibration of the new instrument is recommended prior to switching instruments. [Note that plate and pipette tip materials can cause significant changes in derived results (IC_{50}, EC_{50}). This may be due to changes in the adsorption and wetting properties of the plastic material employed by vendors. Under these conditions a full validation may be required].
- Changes in dilution protocols covering the same concentration range for the concentration–response curves. A bridging study is recommended when dilution protocol changes are required.
- Lot changes of critical reagents such as a new lot of receptor membranes or a new lot of serum antibodies.
- Assay moved to a new laboratory without major changes in instrumentation, using the same reagent lots, same operators (or operators with similar experience and/or training), and assay protocols.
- Assay transfer to an associate or technician within the same laboratory having substantial experience in the assay platform, biology and pharmacology. No other changes are made to the assay.

The purpose of the validation study is to document that the change does not reduce the assay quality.

2.4.2.1. Protocol and Analysis

Conduct the assay comparison portion of the Replicate-Determination Study discussed in Section 2.2, i.e., compare one run of a minimum of 3 to 5 compounds using the existing assay to one run of the assay under the proposed format. If the compound set used in the original validation is available then one need to only run the set again in the new assay protocol, and compare back to Run 1 of the original Replicate-Determination Study. The acceptance criterion is the same as for the assay comparison study: the assay should have Limits of Agreement between -20% and +20% for % activity, and an MSR < 3 and Limits of Agreement should between 0.33 and 3.0 for potency.

2.4.2. Tier II: Substantive Changes

Substantive changes requiring full assay validation: when substantive changes are made in the assay procedures, measured signal responses, target pharmacology and control compound activity values may change significantly. Under these circumstances, the assay should be re-validated according to methods described in Section 2.2. The following changes constitute substantive changes, particularly when multiple changes in factors listed below are involved:

- Changes in strain of animals: e.g., SD to Fischer
- Transfer of the assay to a different laboratory location, with distinctly different instrumentation, QB practices or training.

- Changes in detection instruments with significant difference in the optics and electronics. For example, blood pressure monitors, behavioral test equipment, counting equipment, spectrophotometers, and plate readers.
- Changes in assay platform: e.g.: filter binding to LS/MS detection for ex vivo binding assays.
- Changes in assay reagents (including lot changes and supplier) that produce significant changes in assay response, pharmacology and control activity values. For example, changes in enzyme substrates, isozymes, cell-lines, label types, control compounds, calibration standards, (radiolabel vs. fluorescent label), plates, tips and bead types, major changes in buffer composition and pH, co-factors, metal ions, etc.
- Changes in liquid handling equipment with significant differences in volume dispensing capabilities.
- Changes in liquid handling protocol with significant differences in volume dispensing methods.
- Changes in assay conditions such as shaking, incubation time, or temperature that produce significant change in assay response, pharmacology and control activity values.
- Major changes in dilution protocols involving mixed solvents, number of dilution steps and changes in concentration range for the concentration-response curves.
- Change in analyst/operator running the assay, particularly if new to the job and/or has no experience in running the assay in its current format/assay platform.
- Making more than one of the above-mentioned changes to the assay protocol at any one time.

Substantive changes typically require full validation, i.e., a complete Replicate-Determination Study. If the intent is to report the data in the same AMV then an assay comparison study must be conducted as part of the Replicate-Determination study.

2.5. How to Deal with High Assay Variability

As mentioned previously, the sources of variation in an assay include the between- and within-run sources of variability. In addition, the animal-to-animal variability as well as variability in measuring the response of the subject also contributes to the overall variability in the assay. In order to optimize an assay and/or when an assay fails to meet the acceptance criteria, it is important to specifically assess each of these sources of variability. The variance, which is the square of the standard deviation, can be used to estimate the magnitude, or relative contribution, of each of these sources of variability. Variance is useful because the sources of variation are additive on the variance scale, but not on the standard deviation scale. When an assay fails to meet the acceptance criteria, it is necessary to determine the source of the high assay variability in order to be able to make changes to reduce the relevant variability. For example, simply increasing the number of animals per group may not necessarily reduce between-run variability.

2.5.1. Example: Analyzing Variability in an Ex Vivo Binding Assay

In an *ex vivo* receptor binding assay, subjects (5 rats per dose group) were administered a dose of a test compound orally, sacrificed 1 hour later, the cerebellum removed and stored at -70° until used in the binding assay. For the *ex vivo* receptor binding assay, the tissue was homogenized and incubated for 30 minutes at 37°C in a water bath. The tissue from each animal was aliquoted into 8 tubes, together with radioligand and buffer, and incubated for 2 hours. Four tubes were used to measure total binding and four tubes to measure non-specific binding. Each tube was centrifuged at 12000 x g for 5 minutes, washed and then placed in a gamma counter and counted. Counts for each tube were converted to DPM and recorded. A plot of the data from each of the vehicle treated animals for each of 3 experiments is presented in Figure 8. The variance in the assay is summarized in Table 2.

The total variance was partitioned statistically into study-to-study, animal-to-animal, and tube-to-tube variability. In this example, the result for each of the four tubes is the difference between the DPM in one of the four tubes used to estimate total binding minus the mean of the four tubes used to estimate nonspecific binding. Thus, each of the four tubes represents a different measurement of one subject and the variability among the four

tubes from one animal represents variability in the measurement of binding in that animal. In this example, study-to-study variability accounts for approximately 68% of the total (Table 2, far right column). In order to optimize the assay and reduce the overall variability, reducing the study-to-study variability would have the greatest impact. One way of reducing study-to-study variability is to normalize the data within each study in order to compare results across each study. In this example, the radiolabel was ^{125}I, and its relatively short half-life is the major source of the study-to-study variability; normalizing the data within each study would reduce study-to-study variability due to isotopic decay. The need to normalize, rather than using raw counts, is well known and accepted for *in vitro* binding studies, but this practice is not common for *in vivo* assays. It is appropriate to compare raw signals across studies only when the study-to-study variation is negligible. Note that increasing the sample size would not reduce this source of variability. The second largest source of variability, approximately 18%, is tube-to-tube, that is, the variability in measuring each animal. In order to reduce this source of variability, additional measurements of each animal would be needed. Using 5 tubes to assess specific binding, rather than 4, and only 3 tubes to assess nonspecific binding (which is typically not highly variable) would reduce this source of variability. If only one measurement on an animal is made, the measurement-to-measurement variability still exists, but cannot be calculated. (Another example is that of measuring tumor volume with calipers – measuring only once at a given assessment versus measuring multiple times at a given assessment.) Note that increasing the sample size would not reduce this source of variability. Multiple measurements of the response for one subject are the exception rather than the rule for *in vivo* experiments, and yet variability in the measurement process may be a major source of variance in the study. It is important during assay development and optimization to assess the potential contribution of measurement variability on the total variability in the assay and address this potential source as needed, for example by taking the average (or median) of multiple measures on a given subject. The smallest source of variability in the present example, approximately 13% of the total, was due to animal-to-animal variability. Thus, increasing sample size would have impacted the smallest source of variability. The relative contributions of "animal" and "measurement" variation can be used to determine the optimal number of measurements per animal, as well as number of animals. In conclusion, the relative contribution from different sources of variability needs to be directly assessed during assay development and optimization. This assessment allows one to directly address the most relevant sources of variability in order to optimize and statistically validate an assay.

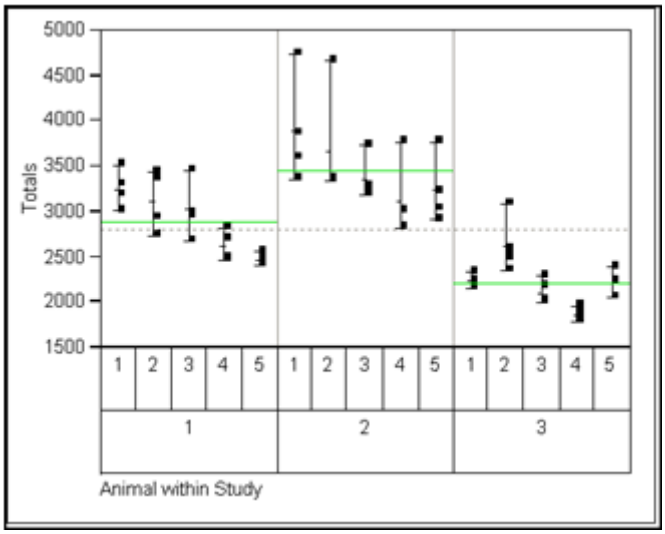


Figure 8: Plot of the vehicle treated animals in an *ex vivo* binding assay. The ordinate (y-axis) is total specific binding in DPM in each of the four tubes used to assess specific binding. Data for 5 animals in each of three studies is represented.

Table 2: Estimated variance contributed by study-to-study variability, animal-to-animal variability, and tube-to-tube variability in an *ex vivo* binding assay.

Source of Variation	Estimated Variance	Estimated Std Dev	Pct of Total (%)
Study	373088	610.8	68.3
Animal	72819	269.9	13.3
Tube	100033	316.3	18.3
Total	545941		100

2.5.2. High Variation in Single Dose Determinations

Table 3 below can be used as a reference to determine the sample size necessary for single dose or dose-response assays with high variability. For a given **coefficient of variation** (CV) of the raw data values based on a sample size of 1 subject, the table shows the number of subjects per dose needed for the CV of a mean to be less than or equal to 10 or 20%.

Increasing sample size to reduce variability will also reduce the capacity (i.e., throughput) of the assay to test compounds. Further optimization of the assay could reduce variability and maintain or increase its capacity. The decision to further optimize or to increase sample size will have to be made for each assay.

Table 3: Sample size necessary to reduce the coefficient of variation (CV) to less than 10 or 20%, given a known CV when the sample size is one.

CV for Individual Subjects	Number of Subjects so that CV Mean < 10%	Number of Subjects so that CV Mean < 20%
<10	1	1
10.1-14.1	2	1
14.2-17.3	3	1
17.4-20.0	4	1
20.1-22.3	5	2
22.4-24.4	6	2

Table 3: continued from previous page.

24.5-26.4	7	2
26.5-28.2	8	2
28.3-30	9	3
30.1-31.6	10	3
31.7-33.1	11	3
33.2-34.6	12	3
34.7-36.0	13	4
36.1-37.4	14	4
37.5-38.7	15	4
38.8-40.0	16	4

2.5.3. High Variation in Dose-Response Determinations

If in Section 2.2 the assay fails either test (MSR > 3 or Limits of Agreement outside the interval 0.33-3) then the variability of the assay is too high for typical purposes. The following options should be considered to reduce the assay variability:

1. Optimizing the assay to lower the variability in the raw data values. Check that the dose range is appropriate for the compound results. Increasing the number of doses and/or subjects per dose may improve the results. A minimum of 6 doses at 2X intervals, and analyzing the data using nonlinear curve-fitting techniques, is recommended. In general, it is better to have more doses rather than more subjects per dose. The doses should cover the expected range of the assay, e.g., 0-100%, as much as possible.
2. Consider increasing sample size as discussed below. Note that the impact of increasing sample size may decrease capacity, and so the Replicate-Determination Study, and a detailed analysis of the sources of variation, should be used to assess whether increasing the number of subjects per dose will achieve the objective.
3. Adopt as part of the standard protocol to re-test compounds. For example, each compound may be tested 2 or more times in different runs of the assay. Averaging the results from multiple runs will reduce the assay variability (**NB**. In such cases the individual run results may be stored in the database and then the data mining/query tools are used to average the results).

To investigate the impact of increasing sample size in the dose-response assay you should conduct the Replicate-Determination Study with the maximum number of subjects contemplated (e.g., 5 subjects / dose). The data can be analyzed first using all available subjects. Then one subject per group can be removed at random and the data re-analyzed. This step is repeated until the smallest sample size is found that still meets the acceptance criteria. An example below will illustrate this idea.

An *in vivo* receptor occupancy assay was run using 1 subject per dose and the Replicate-Determination Study did not meet the acceptance criteria. To examine if replication (i.e., increasing the number of subjects per dose) would help, a new Replicate-Determination Study was conducted using 4 subjects per dose. Table 4 shows the results of fitting ED_{50} curves and re-evaluating the MSR and LsA for 2, 3, or 4 subjects per group:

From Table 4 we can see that it takes all 4 subjects to meet the MSR acceptance criteria, and more than 4 subjects would be needed to meet LsA acceptance criterion. It should be noted that the LsA results are close to being acceptable with 4 subjects per group.

Table 4: Results of fitting ED_{50} curves and re-evaluating the MSR and LsA for 2, 3, or 4 subjects per group.

Subjects	MSR	LsA
2	3.62	0.35 – 4.59
3	3.32	0.43 – 4.74
4	2.44	0.53 – 3.16

3. Design, Sample Size, Randomization, and Analysis Considerations

3.1. Assay (Experimental) Design Considerations

Good experimental design is important in order to answer the research question of interest in a way that is free of bias, can be generalized to the desired or targeted population, and is of sufficient size to properly answer the question. This includes such things as determining the measurements to make, timing, dosing frequency and route, the species to use, etc. It also includes identifying the relevant statistical analyses, determining appropriate sample sizes, and determining a randomization scheme.

Several expectations for good experimental design and analysis include:

- An a priori statement of objectives/hypotheses
- Appropriate experimental design methods
- Sufficient, but not excessive, sample size to ensure statistical significance of biologically relevant effects
- A priori determination of appropriate statistical methods

3.1.1. Objectives and/or Hypotheses

State the objectives and/or the hypotheses you wish to accomplish or test with an assay before beginning an experiment. If a hypothesis is formed after collecting the data, the results may be biased. The objectives should be defined in terms of well defined end-point(s). Examples of objectives include comparison of food consumption between a test compound and control, comparison of survival rates between the treated and untreated groups, comparison of the effect level or the ED_{50} between a test compound and control, etc.

3.1.2. Design Strategy

In vivo studies should be designed in such a way that all meaningful biological effects are statistically significant. In an exploratory study, this "meaningful effect" might correspond to any effect that is pharmacologically relevant. For a project/program team, this meaningful effect might correspond to an effect that meets the CSFs defined in the flow scheme. Power and sample size analysis is especially relevant for assays that are designed to address key endpoints and make decisions as to whether a compound meets the assay's CSF. Biologically meaningful effects are not always well-known in advance, in which case a range of plausible effects could be considered.

To a great extent, experimental design is about identifying and determining a strategy to deal with different kinds of variables. The types of variables encountered in research include:

- Manipulated variable ***(independent/explanatory variable)***
- Response variable ***(dependent/outcome variable)***
- Extraneous variables ***(uncontrolled/random)***

The manipulated variable is a purposeful attempt to introduce variability into the experiment by, for example, administering different doses of a drug. If the manipulated variable were the only source of variability, then research design would be quite simple.

Extraneous variables can encroach upon an experiment and can change the results in ways that we do not want or of which we are unaware. Examples of extraneous variables could include inherent animal variation, time of day, baseline body weights or glucose levels, operator ability or skill, lab noise or activity, etc. **To a great extent, experimental design is about having a strategy to deal with extraneous variables**. To ignore them can all too often lead to biased results and the requirement of larger sample sizes. Fixing (holding them constant) or eliminating them, such as by considering only a subgroup of animals, can reduce bias and sample sizes, but can also reduce the generalizability of the results to only those conditions considered in the experiment. Another approach is to control for them by incorporating them into the experimental design, ideally at the design stage, or at the statistical analysis stage if the former is not possible.

Some additional design considerations include:

- Appropriate random allocation of animals to treatment groups.
- Blinding of observers to drug treatment allocation, whenever possible, especially when subjective evaluations are to be made by observers.
- Proper selection of dose levels.
- Optimal selection of control groups.
- Optimal time points to collect samples.
- Proper statistical methodology.

Different design strategies should be carefully considered to minimize variability and maximize information from the experiment.

The design issues stated above should be addressed in the context of the key endpoints (or summary measures) from the study. Examples of such endpoints may include survival rate, glucose normalization, etc. When there are several endpoints of interest from an assay, certain design questions such as the power of the study should be assessed with respect to the endpoints that are considered to be most important by the scientist and the project team.

3.1.3. Endpoints

The key endpoints from the study must be identified first, as all other design choices should be tailored to these outcomes. Typical outcomes include:

- Statistical significance from control using Analysis of Variance (ANOVA).
- ED_{50} (either absolute or relative) from a dose-response model such as the 4-parameter logistic model (4PL)

Note that the latter can be determined using a variety of methods, and it is important to define the pharmacologically "effective dose," as well.

3.1.4. Control Groups

Control groups serve three purposes: (i) a comparison to the test groups, (ii) as a quality control marker, and (iii) to normalize the response for comparison across studies. A "negative" control is used in all studies and serves as the comparison group. An active or "positive" control is a compound that has a different response from the negative control, and normally represents the maximal response of a standard treatment. A positive control is used when normalization is necessary to stabilize the response across runs of the assay or to illustrate the range of signal available in a particular run of the assay. If a positive control fails to separate from the negative control

then there is an increased chance of a false negative outcome. Since normalized responses may still lack reproducibility across runs, a third control may be employed to monitor the reproducibility of the normalized response across runs of the experiment. This control would be a second positive control, but at an activity level lower than the first positive control. These are used as a quality control marker to establish that each run of the experiment is performing as expected, and hence are often called a "quality control." The activity of a "quality control" should be at a level of activity desired for the advancement of test compounds (see Section 2.3).

3.1.5. Statistical Analysis Plan and Implementation

Before developing and validating an assay to be used on a flow scheme, appropriate statistical methods including data transformations and software for analyzing the data from these experiments should be determined.

There are several statistical methods available to analyze any given experiment/data set, and the choice of these methods and the way a certain class of methods is implemented can significantly impact the conclusions from the experiment. For example, there are certain statistical considerations one should take into account when using the analysis of variance (ANOVA) method, including the distribution of the data, equality of variances, baseline variables, methods for comparing different groups, etc. Also, a two-sample *t*-test might often seem appropriate for several types of experiments, but upon careful examination of the study design, the *t*-test might turn out to be less appropriate than some of the other statistical analysis methods for such experiments.

If the study design is changed at any time during a series of experiments, appropriate analysis methods and implementation strategy should once again be examined in light of these changes.

The basic types of experimental designs are:

- Parallel group
- Randomized block
- Repeated measures
- Cross-over design

3.1.5.1. Parallel Groups Design

In a parallel groups design, subjects are randomly assigned to groups and each group receives one level of a treatment, for example, one dose of a drug, so each group is independent of every other group. This basic design assumes that there are no important *extraneous variables* that we can identify which will influence or bias the results.

Features of a parallel groups design:

- Simplest design.
- Subjects are randomly assigned to groups, and groups are typically, but not necessarily, of equal size.
- Each group receives one level of a treatment, e.g., one dose of a drug
- Use when randomization is possible.
- Does not account for extraneous variables that influence or bias the results. The variation caused by extraneous variables is attributed to the overall assay variability.

3.1.5.2. Randomized Block Design

A second basic design type is called a Randomized Block Design. Randomized block designs are used when an extraneous variable can be identified prior to randomization and subjects can be divided into subgroups based on values of the extraneous variable. Like a parallel groups design, each treatment group is an independent group of subjects. However, subjects are not assigned to treatment groups in an entirely random manner. Rather, subjects are first placed into one of several subgroups based on a blocking or matching factor (such as baseline

values, time of day, gender, baseline body weight, etc.) and then subjects in each block are randomized to the treatment groups.

It is necessary for each subgroup (or "block") to contribute equally to each treatment. Each subgroup must contribute an equal number of subjects to each treatment. Thus, there must be at least as many subjects in each subgroup as there are treatments. For example, if there are 4 treatments, there must be at least 4 or ideally a multiple of 4 subjects in each subgroup.

It is important that a separate randomization be performed on each block so that high or low values of each subgroup are not always placed into one treatment. This strategy forces the extraneous variable to be balanced across treatment groups.

One disadvantage of the randomized block design is that it is not always logistically feasible. For example, the investigator may be aware of an extraneous variable, but not be able to measure it before the randomization process, such as the size of a cardiac or cerebral infarct.

Features of a randomized block design:

- An extraneous variable can be identified and measured before starting the experiment.
- Subjects can be divided into subgroups based on values of the extraneous variable.
- Each subgroup ("block") has as many subjects as there are treatment levels.
- Within each block, treatments are randomly allocated to subjects.
- A separate randomization is performed in each block.
- Forces the extraneous variable to be balanced across treatment groups.
- Not always logistically feasible.

Note: we refer to a factor as a "blocking" factor when it is a continuous measure, for example baseline blood glucose levels, that we can divide into different levels, such as high, medium and low. We refer to a factor as a "stratification" factor when it is not continuous, such as gender and we can stratify the groups on the basis of that parameter.

In a randomized block design, it is necessary for all subjects within a block to be as similar as possible. There are several ways in which blocking can be accomplished:

One of the most common ways to "match" subjects is to rank all of the subjects (e.g., 1 through 20) according to each subject's value of the blocking factor. Subgroups of subjects are then grouped into individual blocks (e.g., subjects 1 – 4 as block 1, subjects 5 – 8 as block 2, etc.). So, within a block, there are similar values of the blocking factor. This minimizes the variance due to the blocking factor ("extraneous variable").

3.1.5.3. Analysis of Covariance (ANCOVA)

Sometimes it is not possible to identify or account for an extraneous variable at the beginning of a study or design phase. In these instances, it may be still be possible to remove the effect of an extraneous variable during the analysis. One technique is to employ an Analysis of Covariance or ANCOVA. Note that if the treatment affects both the response and the covariate then ANCOVA must not be used, because any observed effect may be due to actions of the treatment on the covariate rather than on the response variable.

An ANCOVA model is most useful when there is a linear relationship between the response and the covariate within each group, and when these slopes are similar. If there is sufficient sample size, it is relatively straightforward to test for these conditions. If the slopes aren't different from zero then there is little to no benefit in the ANCOVA model. If the slopes aren't parallel then the interpretation of the treatment comparisons depends on the level of the covariate.

It is also very important that the values of the covariate overlap among the treatments. If the values for the covariate don't overlap, then you are extrapolating into regions where you have no data and the results could be incorrect. Use of proper randomization techniques will usually prevent this situation.

If the experiment was designed as a randomized block design, it is generally best to analyze it as a randomized block design and avoid using ANCOVA.

You may be able to avoid an ANCOVA model by using the baseline to normalize the response (i.e., change from baseline, ratio, etc.; Figure 9).

3.1.5.4. Repeated Measures and Cross-over Designs

Another popular design is to use each subject as a block and test each subject at each of several time points (repeated measures design) or under each treatment condition (cross-over design). With this approach there is only one subject within a block and this minimizes the variance by using each subject as its own control. Repeated measures and crossover designs are just special cases of a randomized block design.

Some examples:

- When a subject receives a dose of drug and is tested at multiple time points.
- When a subject receives all doses of drug. This type of design is also a repeated measures design, but it is a special case of repeated measures and is referred to as a crossover design because each subject is "crossed over" to each treatment or treatment level.
- In a crossover design, different sequences of treatments are identified, and each subject is randomized to one sequence.
- Cross-over designs assume that the effects of each treatment dissipate or don't interfere with the response of the next treatment (i.e., no carryover). If this is not the case, then the cross-over is not an appropriate design.

Repeated measures designs allow us to separate out the variability due to individual differences (that is, to use each individual as their own control) and therefore better evaluate the effects of the manipulated, or independent variable. This increases the power of the analysis and means that fewer subjects are needed to have adequate statistical power.

3.2. Statistical Analysis Considerations

Once the experimental design has been selected, the appropriate statistical analysis then follows:

Parallel groups

- One factor with **only two levels**: t-test or ANCOVA
- One factor with more than two levels: one-way ANOVA or ANCOVA
- Two factors: two-way ANOVA or ANCOVA

Randomized blocks

- One factor: two-way ANOVA
  - Block as the second factor
- Two factors: three-way ANOVA
  - Block as the third factor
- Crossover & repeated measures: two-way and repeated-measures ANOVA
- Stratified designs: two-way ANOVA

The main analysis issues deal with how well the response data match the probability model assumed by the statistical analysis.

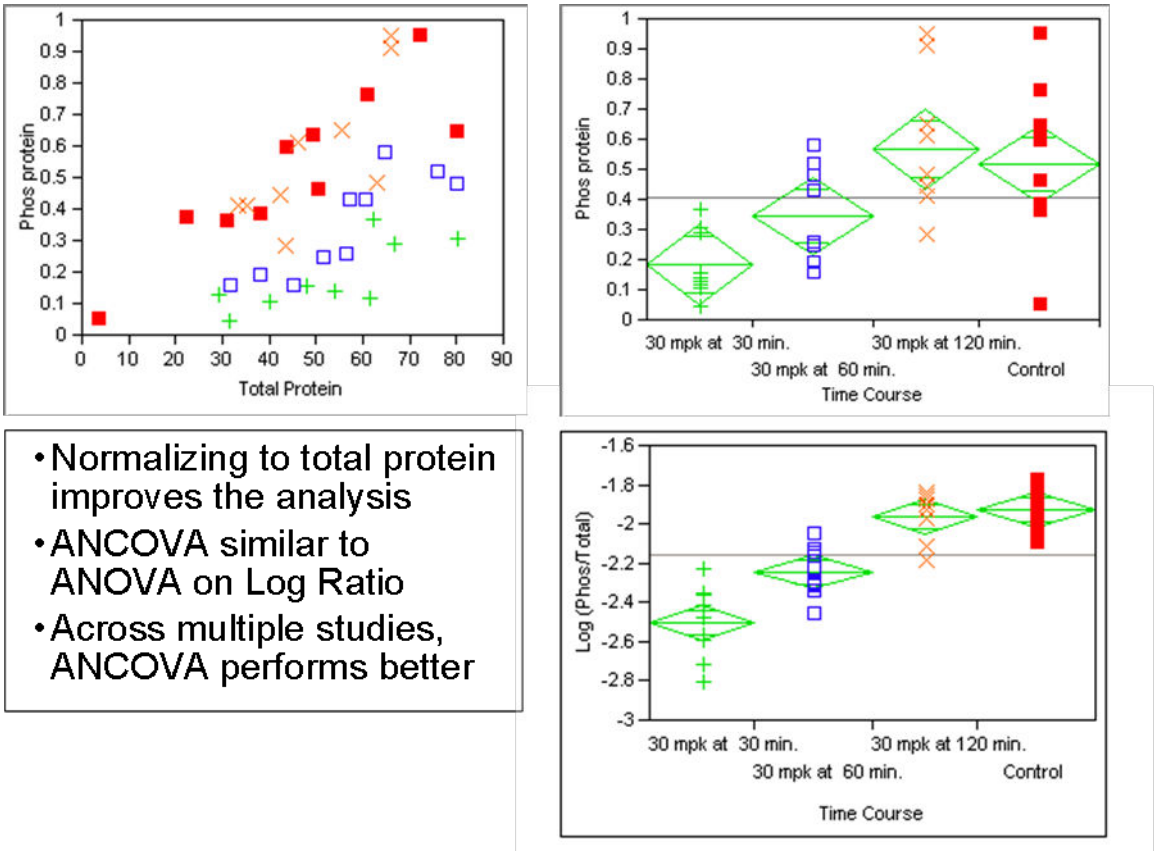


Figure 9: Examples of Analysis of Covariance (ANCOVA).

For ANOVA to identify statistically significant treatments or doses, the main issues are to verify the following.

- The residuals of the response or dependant variable are normally distributed with constant variability across groups
- Techniques to handle outliers are appropriate (see below)
- Appropriate multiple comparison techniques are employed for the study objectives
- If concomitant variables such as baseline measures are used, the analysis should appropriately adjust for these variables

For dose-response studies to estimate an ED_{50}, the main issues to be examined are the following:

- Whether the variability is constant or varies with the magnitude of the mean response
- Suitability of the dose range for the dose-response model
- Suitability of the dose-response model for the biological effects being examined.

For example, to fit a 4 parameter logistic model successfully, a wide dose-range allowing for well-defined tops and bottoms is required. In practice that is difficult to achieve, and often the top or the bottom is fixed at some pre-specified value. This imposes additional biological assumptions that should be assessed.

3.2.1. Outliers

The occurrence of an occasional outlier is not uncommon with *in vivo* experiments, and their presence can sometimes have a large impact on calculated means and, in particular, standard deviations. An observation isolated far above (or below) the bulk of the rest of data can "pull" the calculated mean toward it and result in an overestimate (or underestimate) of the true mean. Since the standard deviation involves a squared term, the impact of an outlier can be even more dramatic, making the standard deviation larger.

In addition to negatively affecting the calculation of summary statistics, outliers can also affect the accuracy of the p-values generated by statistical tests, such as the paired and two-sample t-tests, and the ANOVA. One possible remedy is to perform the statistical test on transformed data (typically the square root or log transform). Transformation is indicated when variability is not constant (e.g., across treatment groups in a one-way ANOVA) and/or when the data are skewed (i.e., longer tail) to the right, but a transformation can also eliminate apparent outliers in some cases (Figure 10).

Another remedial approach is to employ a non-parametric or rank-based statistical test, in which raw data are replaced by their ranks. These are indicated when the data at hand are not adequately modeled by the normal (symmetric, bell-shaped) distribution. Statistical tests based on ranks also down-weight any outliers present in the data. Non-parametric analogs of the paired and two-sample t-tests, and the ANOVA are respectively the Wilcoxon signed rank test, the Mann-Whitney U test, and the Kruskal-Wallis test.

It is acceptable to report the usual summary statistics when presenting data from an *in vivo* experiment for which a transformation or nonparametric test was used for the statistical analysis. However, as mentioned above, outliers can distort the mean and standard deviation, so a better approach is to report summary statistics consistent with the chosen remedial measure (e.g., use the antilog of the mean for log-transformed data; use the median for rank-based tests). Consult a statistician for the appropriate method of reporting summary statistics when outliers are present.

Another approach to dealing with outliers is to simply remove them, but this of course requires one to be able to discern which observations are truly erroneous and which simply represent the underlying variability in the assay. The outlier boxplot is a commonly employed and effective tool for identifying outliers. **However, detection of an outlier by this (or other methods) does not automatically mean the observation can be removed**. It simply identifies observations that have the potential to disrupt the statistical analysis and that should be investigated.

If a valid, assignable cause can be identified for an outlying result, or if the result is simply inconsistent with what is being measured and suggests an error was made, the observation can be removed. Otherwise, the analysis should be performed with and without the outlier(s), with both results reported. **If an observation is removed for cause, it should be documented that the data point existed along with the reason for removing it**. Ad hoc rules based on distance in standard deviations units from the mean should not be used.

Outlier box-plots are shown above the histograms in Figure 10. The box is formed by the first and third quartiles and represents the middle 50% of the data. The length of the box (i.e., third quartile minus first quartile) is the inter-quartile range (IQR). The vertical line within the box is the median, and the horizontal lines connected to the box extend to the extremes of the data that fall within 1.5 times the IQR. Any observations falling outside 1.5 times the IQR from the first or third quartiles appear as points on the box-plot and are potential outliers.

A minimum of 10 observations are recommended to generate a box-plot, but in the context of an experiment comparing several compound activities, each individual compound may be tested in fewer than 10 animals. In that case, the desired statistical model (e.g., a one-way ANOVA) can be fit and the outlier box-plot generated on the residuals from the model fit.

Residuals are automatically calculated in most statistical software, and in some cases are very simple to calculate. For example, the residuals from a one-way ANOVA of compound activities are simply the data "centered" by subtracting the respective compound mean. Consult a statistician for help with generating residuals from statistical models.

Figure 10 is an example of dog plasma exposures for 8 formulations of a single compound (4 dogs per formulation). The residuals from a one-way ANOVA fit to the raw AUCs appear in the box-plot and histogram

in Figure 10A. Note that the histogram suggests right skew for the AUCs, and there are potential outliers in the box-plot, particularly on the upper end of the distribution.

Right skew with outliers (along with increasing variability with increasing mean) are telltale signs that a log transform may be needed. The residuals from a one-way ANOVA fit to log-transformed AUCs appear in Figure 10B. Note that the distribution is much more symmetric, and the box-plot does not identify any outliers on the log scale. The bell-shaped normal distribution (superimposed on the histogram) appears to more adequately model the log-transformed AUCs.

In summary:

- Outlier box-plots can be used to identify potential outliers. Other methods, such as number of standard deviations from the mean, are not recommended.
- An outlier can be removed from an analysis if it has an assignable cause or is clearly erroneous, but this should be documented.
- When there are outliers that do not have a known cause and are not clearly erroneous, analyze the data with and with the suspected outliers and report all results.
- Transforming the response variable can sometimes make outliers become non-outliers and satisfy standard analysis assumptions better than non-transformed data.
- Non-parametric analysis methods can be used to minimize the impact of outliers on analysis results without removing them from the analysis.

3.3. Randomization

There are a number of randomization techniques that are available. Certain study designs require specific randomization techniques, (e.g. randomized block designs, stratified designs, etc.). Random numbers should be obtained from an acceptable random number generator, (e.g., Excel, JMP, SAS, random number table, randomization web tool, etc.).

It is also very important to appropriately randomize subjects to treatment groups, as this reduces opportunities for bias. Randomization requires extra time and effort, but it can be more costly to not use it. Non-random strategies for assigning animals to treatment groups, when applied consistently across studies, will tend to introduce flaws, or bias, into the study results. By looking at the performance of a strategy across a series of studies we can then examine how well or poor a particular strategy is working. A truly random selection process is one where each subject is equally likely to be selected for any treatment group. The word "random" has a specific meaning that implies a formal random process such as a random number table or a computer generated random list has been used to make animal selections. **Using a computer-generated random list or method is the preferred method.**

Intentionally assigning subjects to different groups in order to balance one or more baseline variables is not an acceptable randomization method. While it might appear that the goals of randomization have been achieved, other unknown biases could be introduced.

- Randomization:
- Prevents implicit/explicit bias in treatment assignment
    - Approximately balances groups with respect to all extraneous variables affecting response
    - Does add to logistical complexity
    - Is the primary technique for assigning animals to groups
- Work with a statistician to
    - Design a workable randomization
    - Incorporate known covariate information

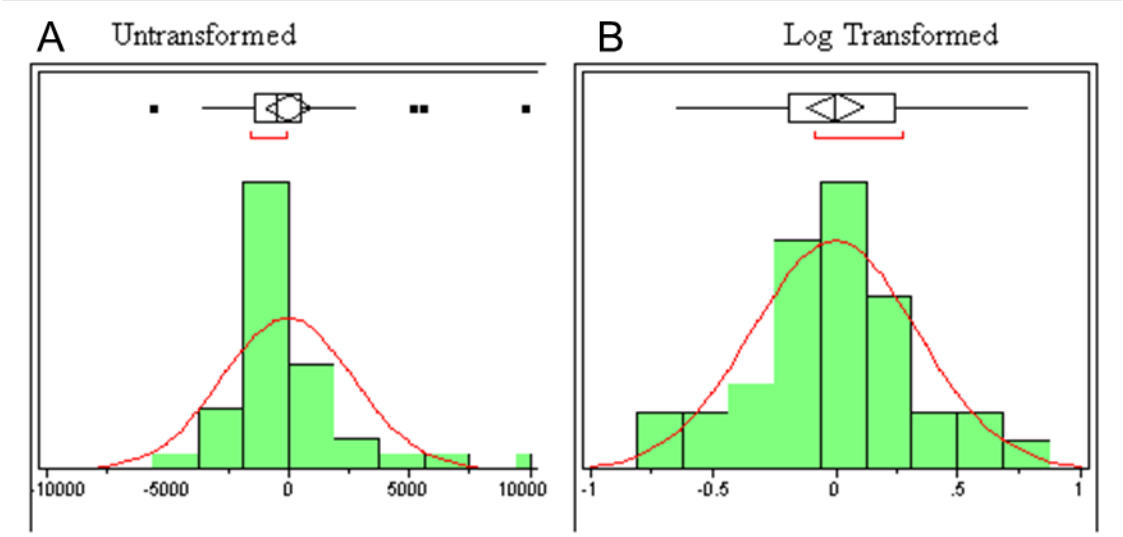


Figure 10: Results of a transformation eliminating outliers. (A) The residuals from a one-way ANOVA fit to the raw AUCs. (B) The residuals from a one-way ANOVA fit to log-transformed AUCs.

3.4. Power

The study should be planned so that the minimum pharmacologically relevant change or the flow scheme CSF has a high probability (usually 80%) of being statistically significant. The probability that a relevant change will be statistically significant is referred to as statistical power.

Note that since the total number of animals available for an assay run may be constrained by practical considerations such as a processing capacity, etc., the power analysis often determines how many dose groups may be examined within a single study.

The statistical power is a function of the following key elements.

- Assay variability (the lower the assay variability the higher the statistical power for a given sample size and effect size)
- Effect size (the larger the effect size the larger the statistical power for a given sample size and level of assay variability)
- Total number of animals in the test group
- Total number of groups and the number of animals per group in the study (in both cases the more animals the higher the statistical power).

Standard convention is to use 80% power to detect the minimum biologically significant effect with a false positive rate of 5% when evaluating or setting sample size. Declaring statistical significance when p-values are less than 0.05, along with appropriate multiplicity corrections, ensures the false positive rate is 5%. In general, a minimum of three runs of an assay with the format to be used in production should be used to estimate the experimental error. The power calculation should account for multiplicity corrections. The number of animals used should be sufficient so that all relevant drug effects are statistically significant.

When setting CSFs for an *in vivo* study, the CSF should be set above the minimum detectable difference if the CSF is defined on the response scale (e.g. percent inhibition > x% at y dose, body weight change > x grams, etc.). For potency CSFs, the study should be powered to exceed the minimum biologically significant effect. Following this paradigm, it is unlikely that you will have sufficient sample size to declare statistical significance at the

minimum biologically significant effect, but you should power the study to be able to detect statistical significance at some dose.

In a dose-response study to estimate an ED_{50}, two compounds will have statistically different ED_{50}s if the ratio of the ED_{50}s (larger::smaller) exceeds the minimum significant ratio (MSR) of the assay (see Section 2.2). Thus, the latter should be small enough to discriminate all pharmacologically relevant differences. The MSR depends upon the number of animals, the number of concentrations used, and the spacing of the doses with respect to the ED_{50}. The number of doses should either be large enough to estimate an ED_{50} over a reasonably wide range and/or adjusted for each compound based on the efficacy demonstrated in a single dose screen. If a large number of doses are used, the number of animals per dose may be quite small.

3.5. Analysis of Dose-Response Curves: Principles and Practice

In considering how to statistically evaluate dose-response curves, it is informative to review some of the historical context of how dose-response curves have been analyzed. When pharmacologists were first determining dose-response curves, in the very early days of pharmacology, they often plotted the data on an arithmetic scale (Figure 11A), that being the simplest and, perhaps, most familiar, to them.

In order to summarize and compare different dose-response curves, scientists attempted to describe these curves mathematically. However, the curves, a type of parabola, require rather complicated equations to describe. Prior to the invention of calculators and computers, it was too laborious and time-consuming to solve these equations. Scientists therefore searched for other ways of plotting and summarizing their data which might be readily described by equations. One way of doing this was to plot the dose-response curve on a semi-logarithmic plot, as shown for the same data in Figure 11B.

The semi-log plot of the data is the familiar sigmoidal (S-shaped) plot of many dose-response curves. This type of plot had the important advantage that **the middle part of the curve was approximately linear** (Figure 12). The portion of the curve from 16% efficacy to 84% efficacy can be described by a linear equation ($y = mx + b$) which can be solved without the aid of calculators or computers. However, data outside of this range (below 16% and above 84%) was off the linear portion of the curve and therefore was excluded from the analysis since the inclusion of such data would alter the slope of the line.

Pharmacologists wanted to have a single number to describe the dose-response plots of their data. For this purpose, the dose which produced a 50% effect (ED_{50}) seemed ideal since it was in the middle of the linear portion of the curve and could be calculated from the linear regression. These types of considerations led pharmacologists to design experiments which emphasized the middle, linear portion of the curve. Pharmacologists therefore primarily designed studies to have three groups (the minimum number of points needed to describe a line), and to have large numbers of animals in each group (to have a robust estimate of each mean).

A problem with this approach was that if the means of any groups fell outside of the 16 to 84% range, they were of little use in solving the linear equation. Many text books taught that data outside of the 16 – 84% range should be excluded, particularly the results of any groups where the mean was 0% or 100%. However, data at the extremes, or asymptotes, were very important data as they defined the top and bottom of dose-response curves. Moreover, it was difficult to calculate confidence limits on the slope and ED_{50}. Thus, using only linear regression and the linear portion of the curve had substantial limitations.

Since sigmoidal curves are nonlinear, a non-linear regression algorithm should be used to fit the data. Today, with computers to solve complex equations very rapidly, we can use non-linear curve fitting techniques to model, or mathematically describe sigmoidal-shaped dose-response curves. From the non-linear curve-fit, specific parameters are estimated which describe the dose-response curve. The parameter estimates can then be used to compare dose-response curves for different compounds.

Sigmoidal dose-response curves can be described using four different parameters:

- The top, or maximum effect
- The bottom, or minimum effect
- The ED_{50}, or mid-point on the curve
- The slope of the curve

Using four parameters to describe a non-linear curve is called a 4 parameter logistic model, as illustrated in Figure 13.

The 4-Parameter Logistic (4PL) Model generates a family of curves using the four parameters of top, bottom, middle and slope. With these parameters, and using nonlinear regression, we can describe most sigmoidal curves. On occasion, there may be a practical or theoretical reason to define what the bottom and/or the top of a curve will be; for example, it may be known from the experimental methods or pharmacological theory that the bottom will be 0% or the top will be 100%. In such cases, only three parameters may be needed to describe the data. In such a case, a 3-parameter logistic model may be used to describe the data. Three parameter models are used most commonly when the dose range is imperfect (too high or too low) with respect to the potency of the compound, and we do not have doses which yield data near the top or bottom.

One also could fix both the top and bottom at constant values (a 2-parameter logistic model). However this approach makes some strong assumptions and one should let the data estimate the top and bottom whenever possible.

3.5.1. Analyzing Dose-Response Data

One way of analyzing data from dose-response determinations is to use ANOVA with a Dunnett's test to determine if there is an overall effect of the drug, and which doses produced an effect was statistically different from the control group. However, this analysis doesn't allow one to estimate a dose producing a 50% effect nor estimate the minimal and maximal effects or the slope. The dose producing a 50% effect will typically lie between two of the doses tested, and therefore requires interpolation to estimate. Interpolation requires some type of regression analysis.

In certain situations, for example with a limited number of dose groups, linear regression may be the best approach. However, the linear regression approach has several limitations:

1. Interpolating between two doses does not use all the data and hence can be inefficient. If one uses more than two doses, a straight-line fit may not be appropriate and the result will be a distorted (biased) estimate of the ED_{50}.
2. It is not trivial to quantify the precision (i.e. calculate a standard error) of the estimate of the ED_{50} from linear regression. Thus, it is not trivial to determine if two ED_{50} values are statistically significantly different.
3. There is no ability to identify the minimum and maximum effect, nor the precision of the estimates, with linear regression.

By using a nonlinear regression model, one can get a better fit to **all** of the data, as well as calculate estimates and 95% confidence limits for the ED_{50}, slope, maximum effect and minimum effect.

In the process of assay validation, methods for analyzing dose-response data must be selected and the method should remain constant for a validated assay. However, there may be situations where deviations are appropriate. Consult with a statistician for how to most appropriately deal with specific situations, or datasets, where deviations from pre-determined data analysis approaches may be appropriate.

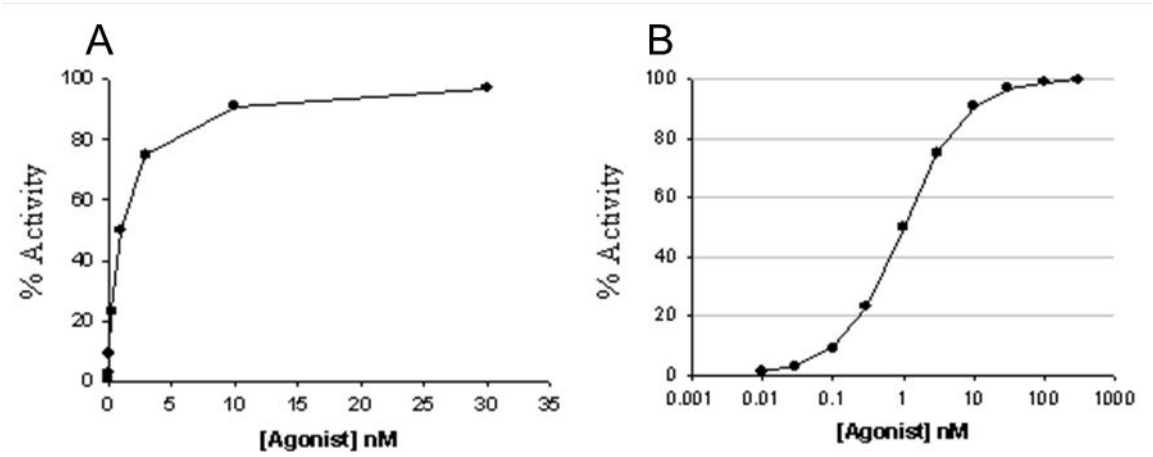


Figure 11: (A) Linear versus (B) semi-logarithmic plots of the same set of theoretical data.

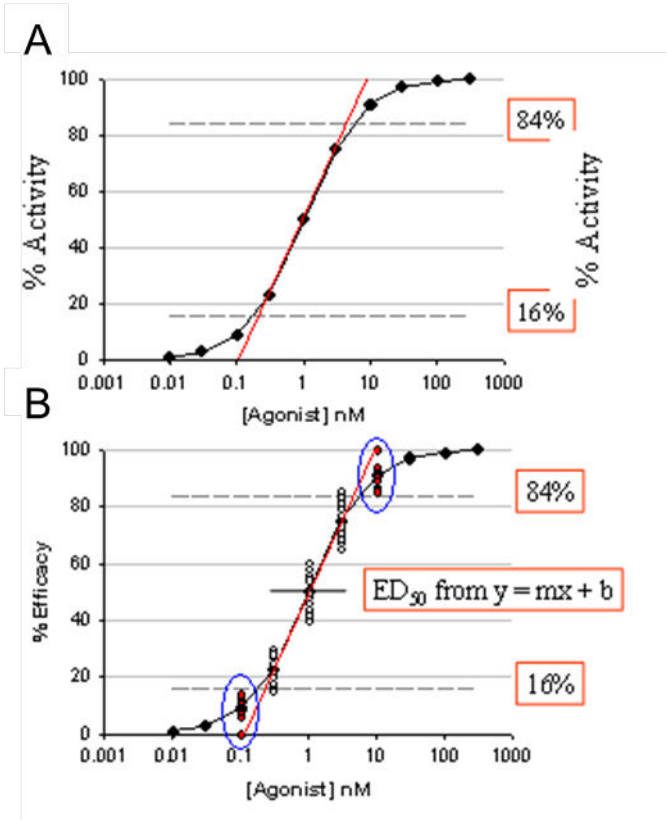


Figure 12: (A) Illustration of the linear portion of a sigmoidal dose-response curve and (B) the influence of data points which lie outside of the 16 – 84% region of the dose-response curve on the calculation of the slope of the line and potentially on the dose producing a 50% effect (ED_{50}).

3.5.2. Experimental design requirements for linear vs. nonlinear analysis

For analysis by linear regression or ANOVA followed by a Dunnett's test, 3 or 4 groups of usually 5 to 10 animals per group traditionally has been required to provide precise estimates of the mean and variation for each treatment group and thus, sufficient power to identify all important treatment effects. To estimate potency results, such as ED_{50}, the requirements are quite different; having data over the entire range of the dose-response curve is required, and estimates of the mean at each individual dose do not need to be as precise. That is, with

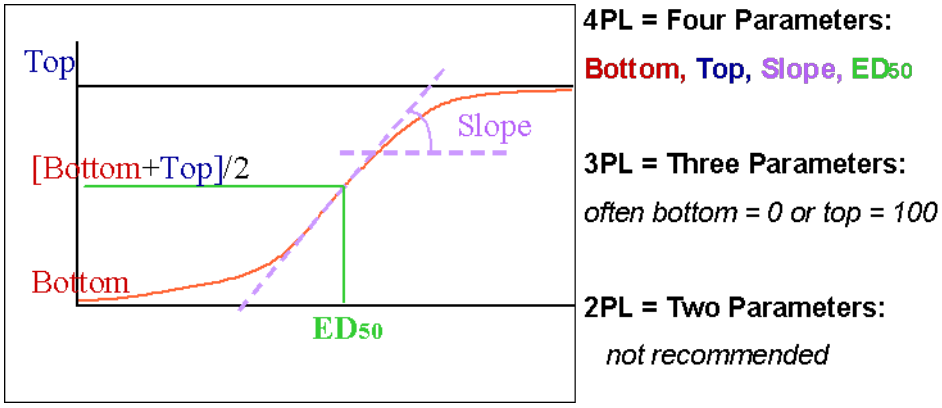


Figure 13: Illustration of the parameters of a 4-parameter logistic model of sigmoidal-shaped curves.

nonlinear regression, one is not estimating a given point (or mean), but rather the parameters of the entire curve; modest deviations in a single point are less likely to substantially impact the parameters of the overall curve. On the other hand, not having data points at, for example, the top or bottom, can substantially impact the analysis. Therefore, it is necessary to have a larger number of groups with fewer animals per group. With experimental designs for nonlinear analysis, it may even be possible to reduce the total number of animals needed. Diagnostic checking of the resulting curve fit is essential; the ED_{50} and asymptotes must make sense. Further, the confidence limits should be sufficiently precise to meet research objectives. Therefore, it is critically important to carefully consider the selection of doses in an experimental design with nonlinear regression analysis. For example, using 6 dose levels with 4 animals per group would be far more appropriate than using 3 dose levels with 8 animals per group. Also for nonlinear regression analysis, it is best to have doses that are equally-spaced on the log scale. *When an assay is statistically validated, the number of groups and animals per group needed should be verified with a statistician.*

As an illustration, the effects of a drug on decreasing ethanol consumption in a behavioral assay were compared under two different experimental designs. For these experiments, animals were trained to drink an ethanol solution over a period of several weeks. Animals were then randomized to treatment groups which receive either vehicle or a dose of the test drug. The drug produced a dose-related decrease in ethanol consumption (Figure 14). In the more traditional experimental design (Figure 14A), vehicle and 4 doses of drug were administered to groups of 6 animals each, for a total of 30 animals. When the data were analyzed by ANOVA and a Dunnett's test, doses of 3, 10 and 30 were significantly different from vehicle. There were only 2 doses on the linear portion of the dose-response curve (16 – 84%), an insufficient number to properly use linear regression, and therefore an ED_{50} and confidence limits could not be calculated with any degree of robustness. If nonlinear regression analysis is applied to the data in the left panel, an ED_{50} value of 0.6 mg/kg is obtained (a value which appears to be too low based on visual inspection of the data) and confidence limits can be obtained only for the parameter Top. The bottom could be fixed at zero, which provides a better estimate of the ED_{50}, but its lower confidence limit could not be calculated using JMP software. When a nonlinear-compatible design was used (Figure 14B), 6 doses of drug were administered to groups of 3 animals each for a total of 18 animals. A low, inactive dose of drug was used in place of vehicle, allowing for a better estimate of Bottom in this experiment. A vehicle group could also have been included. (Note: results from a vehicle treated group can, if appropriate, be incorporated into the nonlinear analysis; however, since there is no "zero" point on a log scale, vehicle is typically assigned a dose-value two to three orders of magnitude below the lowest dose tested; consult with a statistician for ways to incorporate a vehicle group into a nonlinear analysis). From nonlinear regression, an ED_{50} value of 1.5 mg/kg is obtained (which appears reasonable based on visual inspection of the data) as well as 95% confidence limits (0.91 to 2.4 mg/kg). In addition, estimates of the Top, Bottom and Slope, along with their respective confidence

limits, are obtained. Thus, if the primary goal is to obtain an ED_{50} value together with confidence limits, a nonlinear-compatible design yields more results with far greater precision, and may also require fewer animals.

3.5.3. Key points in the analysis of dose-response curves

Analysis Key Points:

- Regression is needed for interpolation
- 4 parameter logistic model is primary model for dose-response work
- Diagnostic checking is essential
  - Reasonable asymptotes
  - Numerically complete answer
  - Try fixing top or bottom as necessary

Design Key Points:

- Doses need to be spaced across a broad dose range
  - Doses may need to be in 2X steps rather than 3X (i.e., half-log)
- Preferable to use more groups but may need fewer animals per group
  - Nonlinear approaches can provide more information with fewer animals

4. Abbreviations

3PL, 3-parameter logistic

4PL, 4-parameter logistic

AMV, Assay method version

CSF, Critical success factor

DLs, Difference Limits

DRC, Dose-response curve

ED_{50}, Dose which produces 50% effect (Effective Dose 50%)

Relative ED50: Dose which produces 50% of the maximal response produced by that drug in the test system; considered to be relative measure of affinity

Absolute ED50: Dose which produces 50% of the maximal response which can be observed in the test system by a positive control; theoretically a drug could produce an effect greater than the maximum that can be measured

GM, Geometric Mean

LsA, Limits of Agreement

LsAd, Limits of Agreement on differences

MD, Mean Difference

MR, Mean Ratio

MSD, Minimum Significant Difference

MSR, Minimum Significant Ratio

RLs, Ratio Limits

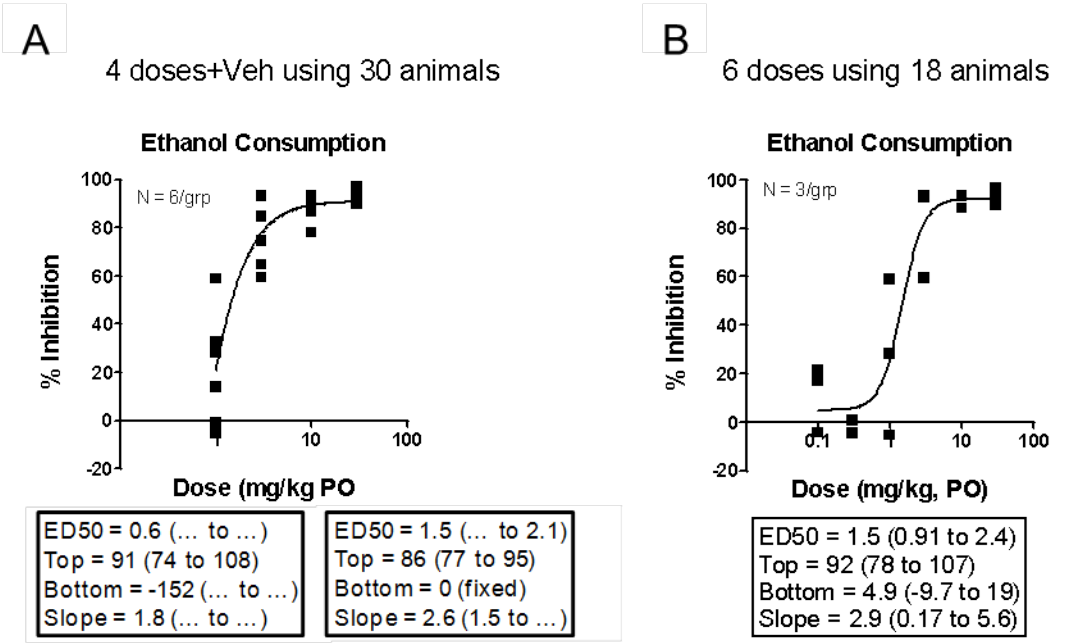


Figure 14: Comparison of (A) results with a more traditional experimental design using vehicle plus 4 doses of drug and 6 animals per dose and (B) results from a nonlinear-compatible design with 6 doses of drug and 3 animals per dose.

SAR, Structure-activity relationship

SDS, Single-dose screen

5. Suggested Reading

Landis SC, Amara SG, Asadullah K.et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012;490(7419):187–91. PubMed PMID: 23060188.

License