Report

On

STATISTICAL METHODS FOR EVALUATING VARIABILITY IN AND SETTING UP PERFORMANCE CRITERIA FOR RECEPTOR BINDING ASSAYS

July 9, 2007

Report for:

Mr. Jim Kariya and Dr. Shirlee Tan
U.S. Environmental Protection Agency
Endocrine Disruptor Screening Program
Washington, D.C.

Prepared by

Dr. Yutaka Aoki
Consultant
DUNS # 788730278

Table of Contents

List of Tables

List of Figures

Selected definitions and abbreviations

Replicates: A set of measurements carried out at a single concentration of an analyte, close in time and space (e.g., adjacent wells), using common biological materials, reagent, and instrumental setting. In the binding assay data to be dealt with in this report, replicates are % binding values usually measured in triplicate.

Run: A set of replicates for an analyte(s) over a range of concentrations performed together on a single occasion, often with accompanying measurements concurrently performed for a reference chemical and positive control. Runs are often performed by different technicians and/or on different days.

SE: standard error

LS: least square

Executive Summary

EPA is validating receptor binding assays to be used to screen chemicals for their endocrine disrupting potential. One of the goals in the validation is to describe within-laboratory and between-laboratory variabilities of the assays. It is necessary to determine an acceptable level of variability for these assays, and then to demonstrate that laboratories obtain results that are within the acceptable level.

In order to ensure low levels of these variabilities, EPA developed performance criteria for a rat uterine cytosol estrogen receptor (RUC ER) binding assay. This document provides a detailed description of the methods for deriving such criteria so that a new set of performance criteria for another assay (or for the same assay based on a new set of interlaboratory data) may be developed in the future.

These methods are based on descriptions of within-run, between-run, within-lab, and between-lab variations of the estimates of four parameters that describe a chemical's interaction with a receptor of interest. The conventional approach for describing these variations was deemed suboptimal, and a new approach was sought and developed.

The interaction between a test chemical and receptor is described by a sigmoidal curve showing how much the test chemical displaces a reference chemical as the concentration of the test chemical increases. The sigmoidal curve is represented by four parameters, i.e., top plateau level, bottom plateau level, $\log IC_{50}$ (or $\log EC_{50}$) and Hill slope. The relative binding affinity (RBA) of a test chemical for the receptor (relative to the binding affinity of the reference chemical) is expressed as logRBA, which is the difference between the $\log IC_{50}$ for the test chemical and that for the reference chemical. Variation of individual binding level measurements also is of interest.

The performance criteria that were developed may be separated into two groups, accuracy criteria and precision criteria. The accuracy criteria consist of upper and lower limits for top, bottom, $\log IC_{50}$, Hill slope, and logRBA estimates. The precision criteria include upper limits for within-lab variation of $\log IC_{50}$ and logRBA estimates as well as within-run variation of individual binding level measurements.[1]

[1] In response to feedback from Dr. Feder (Feder, 2007a, 2007b, 2007c) on an earlier draft of this report (Aoki, 2007a), EDSP decided to substantially revise performance criteria numbers in this document. The revision was to be performed by Data Coordination Center at Battelle. Some of the methods described in this report were to be used in the revision while some other methods were to be replaced by improved counterparts chosen by DCC after discussion with EDSP. This report was prepared as a part of contract awarded to Yutaka Aoki working as an independent contractor. The original purpose of preparing this report was to document his statistical work regarding data from receptor binding assay during his tenure as an ASPH Fellow at EDSP. During its preparation, certain needs for additional work were identified, and limited amount of additional analyses was performed to enhance the integrity of the report. Dr. Feder's feedback was very thorough and helpful to EDSP's efforts to produce justifiable performance criteria. He identified items in the draft report that require further improvement. In order to preserve the original scope of the contract to the extent possible, the revision based on feedback from Dr. Feder was limited to at minimum per EDSP's request. As such, some necessary corrections were made on the text, formulae, etc., in the description of methods in Chapters 1 and 2, but the most of numerical results in Chapter 3 generated

using the originally developed methods were left unrevised. A separate document (Aoki, 2007b) provides responses to main comments in Feder (2007a and 2007c).

1 Introduction

1.1 Overview

The Endocrine Disruptor Screening Program is developing and validating assays for detecting the potential of a chemical to interact with the endocrine system. A group of *in vitro* assays under consideration are called receptor binding assays and have the following common features.

- A radioactively-labeled natural ligand is allowed to bind to a receptor
- A non-labeled test chemical is then introduced at increasing concentrations to see how much it takes to displace the radioactive ligand

Whether the displacement occurs at all is the primary outcome to determine. If it does, the concentration of the test chemical at which a pre-specified level of displacement occurs is estimated as a summary measure of the test chemical's capacity to displace the radioactive ligand, which may be interpreted as a measure of the test chemical's potential to interact with the receptor.

Estrogen receptor (ER) binding assays and androgen receptor (AR) binding assays belong to this group. In order to improve precision of the above-mentioned summary measure, it is customary to run three replicates side-by-side and combine the results, and to perform the experiment three times on different occasions (i.e., perform separate "runs") rather than to perform a single run without replicates.

EPA needs to assure that results from one laboratory are comparable to results from other laboratories. For this reason, it intends to place limits on the quality of data collected for certain standard[2] and positive control compounds that will be run every time the assay is performed. In general, we would like the summary measure to be within a reasonable range and to have low variability; and for the parameters that describe the shape of the concentration-binding relationship to be within biologically- and experimentally-plausible ranges. In order to derive the various limits to be imposed, a proper method for summarizing data across runs within a lab and then across labs by describing data variabilities is needed.

A conventional way of summarizing data across runs (and across labs) is to take a simple algebraic mean of run-specific (lab-specific) parameter estimate and evaluate its variability based solely on the observed variability of such means. It falls short because it does not explicitly distinguish within-run variability from between-run variability, or within-lab variability from between-lab variability.[3] For instance, if results from multiple laboratories look like they might be different from one another, it would not be clear whether they are truly different from one another, or whether they appear to vary simply

[2] The terms "standard chemical", "standard compound", etc., are used in this report. They are more appropriately termed "reference chemical", "reference compound", etc.

[3] Dr. Feder (Feder, 2007b) disagrees with this and state "the conventional method falls short because it does not account for correlation within runs or labs." It seems the disagreement is at a semantic level.

because within-lab variation is high. Also, various published assay validation guidelines require separate estimations of within-lab and between-lab variabilities.

EPA has developed a method that properly analyzes the two sources of variability (within-unit and between-unit) and established performance criteria by applying the method to a set of historical data. The performance criteria will be used in an upcoming interlaboratory validation study.[4]

1.2 Statistical Background

1.2.1 Raw data

A typical data set from a single run of a receptor binding experiment (competitive binding assay) is shown below.

x: log(concentration of the test chemical, M) [5]
y: Percent binding of indicator ligand (i.e., radioligand) to the receptor

x	y
-7.0	0.8
-7.0	0.6
-7.0	0.0
-8.0	11.6
-8.0	12.9
-8.0	13.1
-9.0	59.7
-9.0	55.9
-9.0	57.6
-9.5	79.6
-9.5	78.8
-9.5	85.8
-10.0	98.1
-10.0	86.5
-10.0	104.4
-10.5	104.3
-10.5	96.8
-10.5	99.6
-11.0	104.9
-11.0	103.3
-11.0	101.7

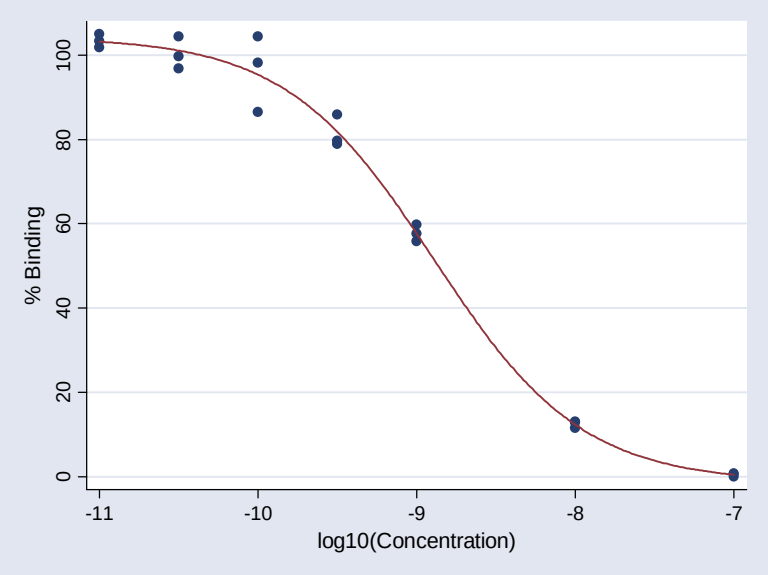

Figure 1.1 Typical data from a receptor binding experiment

Data from a competitive

[4] As will be discussed later, decisions to not impose some of the derived criteria for the ongoing interlaboratory study have been made. This document focuses more on explaining approaches to quantitatively describe data variability and use to such quantitative information to derive performance criteria rather than which criteria are useful in the light of the agency's need and therefore should be imposed. Some discussion on merits and demerits of subsets of the derived performance criteria will be included in this report such that the discussion would be useful for the EDSP when it makes further decisions as to which performance criteria should be developed and enforced for the ongoing and future interlaboratory studies.

[5] In this document "log" is used to mean "logarithm of base 10" unless otherwise noted. This applies to abbreviations such as $\log IC_{50}$, $\log EC_{50}$, and logRBA.

binding assay usually are plotted in the format shown above. The curve shown in red is a fitted curve to be explained below. For standard chemicals, a pre-specified set of concentration levels are used, which typically cover about 4 units in logarithm of base 10.

1.2.2 Model and parameters

In order to describe the data, a model is fit to the data. A common choice of model is the 4-parameter Hill equation model:

$$Y = B + \frac{T - B}{1 + 10^{(\mu' - X)*H}} \quad \text{.....1.1}$$

where μ' = logEC$_{50}$, H = Hill slope, T= top plateau level, B = bottom plateau level. This model is consistent with the law of mass action under the assumption that the labeled and unlabeled ligands compete for a single binding site (Motulsky and Christopoulos, 2003). For the reasons to be explained in Chapter 2, we use the following equation, a modified version of the standard equation:

$$Y = B + \frac{T - B}{1 + 10^{[(\mu - X)*H + \log(\frac{T-B}{50-B} - 1)]}} \quad \text{.....1.2}$$

where μ = logIC$_{50}$.

The parameter of primary interest is logIC$_{50}$, which is defined as x at which y is 50%. On the other hand, logEC$_{50}$ is x at which y = (B+T)/2, i.e., the midpoint between the top and bottom of the curve. The difference between these two parameters is illustrated in Figure 1.2.

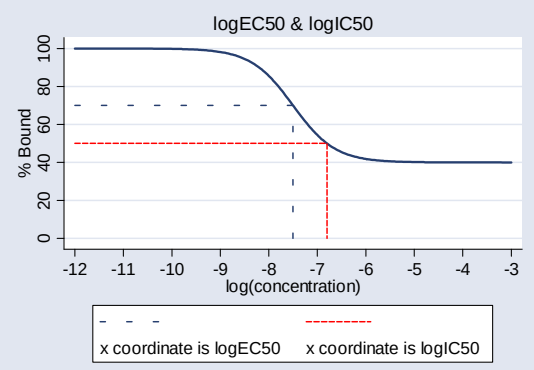

Figure 1.2 logEC$_{50}$ and logIC$_{50}$

The four parameters are estimated using a nonlinear regression to be discussed in detail later. It should be noted that the collected data do not always allow estimation of $\hat{\mu}$ because sometimes the underlying curve does not cross a horizontal line at y = 50%. However, even for such cases, logEC$_{50}$ may be estimated.[6]

[6] Note that logIC$_{50}$ and logEC$_{50}$ each has a unit of log(concentration in moles/liter), and there are counterparts for these expressed in a unit of absolute concentration, i.e., IC$_{50}$ and EC$_{50}$. For the reasons explained in Chapter 2, the use of IC$_{50}$ and EC$_{50}$ when evaluating variability and setting up performance criteria is to be generally avoided. We use logarithms of base 10 of IC$_{50}$ (or EC$_{50}$) instead.

1.3 Statement of the problem

1.3.1 Need for accuracy and precision

EPA is considering defining a binder in terms of bottom plateau level and/or $\log IC_{50}$ for each chemical. In addition, $\log IC_{50}$ may be used as a measure of binding affinity (relative to a standard whose IC_{50} is also measured). Both of these need to be estimated with sufficient accuracy and precision. Note that these terms, "accuracy" and "precision", are used with a specific statistical meaning as described below. For a given true value of a parameter, e.g., μ, accuracy has to do with $E(\hat{\mu}) - \mu$ (i.e., "expectation of $\hat{\mu}$" minus μ or its true value) or bias, and precision has to do with standard error of ($\hat{\mu}$) where $\hat{\mu}$ is an estimate of μ.[7] An ideal estimate would be both very accurate and precise.

The $\log IC_{50}$ estimates may vary widely across chemicals, but for well-behaved chemicals such as standards, the estimates of the other three parameters are supposed to be within a relatively narrow range. Theoretically, a true one-site competitive binder should produce data consistent with a top plateau of 100 %, bottom plateau of 0%, and Hill slope of -1. Deviation from these biologically plausible values for standards indicates problems regarding the data and renders the data suspicious. Thus a check for these parameters would be useful in ensuring data quality, but this can only be done using well-characterized standard and control chemicals. The curve fitted to data on estradiol, a natural ligand of the ER receptor, should have parameter estimates close to these default values if experiments are performed properly. Estradiol is a standard compound for the ER binding assay and will be tested concurrently with any test chemical. Sufficiently high accuracy of estimates of these three parameters for estradiol should be demonstrated by laboratories that participate in the screening of chemicals.

A summary measure called the logarithm of the relative binding affinity (logRBA) may be estimated for a test chemical. logRBA is defined as follows:

logRBA = ($\log IC_{50}$ for estradiol) - ($\log IC_{50}$ for test chemical)

A large logRBA value for a chemical indicates it has strong affinity with a receptor relative to the standard. Performance criteria for logRBA can be established for a specific positive control, e.g., norethynodrel, relative to estradiol for the estrogen receptor binding assay. Criteria for logRBA of a positive control might ultimately turn out to be more useful than that for $\log IC_{50}$ since logRBA may be used as a primary measure of binding affinity for chemicals of unknown binding affinity. LogRBA would be unchanged when there is a parallel shift in $\log IC_{50}$ for the standard and that for the positive control as may occur when there are differences in dilution and other techniques that would be consistent within one lab, but differ between labs.

Discussions so far have dealt with EPA's need for accuracy of parameter estimates. EPA also needs to ensure precision of key parameter estimates. In other words, we would like

[7] In general $\hat{\theta}$ is used as shorthand for an estimate of parameter θ and is read as "theta hat".

$SE(\hat{\mu})$ to be small. Specifically, precision of $\log IC_{50}$s for the standard and test chemicals is of EPA's interest in that the precision of logRBA of an unknown, which is more meaningful than the $\log IC_{50}$ for the unknown, is largely determined by precision of the $\log IC_{50}$s. EPA also considered imposing a criterion for precision for replicates (i.e., an upper limit for variation in individual replicates measured at the same concentration in a single run), which we call within-replicate-set variability. This measure of variability can be estimated after each run, rather than after three runs, it would have use as a preliminary checkpoint for data variability. Imposing a limit for within-replicate-set variability potentially is a simple and efficient way to ensure data quality. We could also impose limits for between-replicate-set variability in y given x and total variability in y given x.

To summarize, the need for the following components of performance criteria was recognized and procedures for deriving them were developed.

Table 1.1 Components of performance criteria

Parameter	Accuracy criterion	Precision criterion
$\log IC_{50}$ (μ)	$\mu_{\min} < \hat{\mu} < \mu_{\max}$	$\log(SE(\hat{\mu}))$ < Max
Top (T)	$T_{\min} < \hat{T} < T_{\max}$	Not derived
Bottom (B)	$B_{\min} < \hat{B} < B_{\max}$	Not derived
Slope (H)	$H_{\min} < \hat{H} < H_{\max}$	Not derived
logRBA (ρ)	$\rho_{\min} < \hat{\rho} < \rho_{\max}$	$\log(SE(\hat{\rho}))$ < Max *
Within-replicate-set variation of y	Not applicable	$SD_{\text{within-replicate-set}}(Y)$ < Max
Between-replicate-set variation of y given x	Not applicable	$SD_{\text{between-replicate-set}}(Y)$ < Max
Total variation of y given x	Not applicable	$SD_{\text{total-between-replicate-set}}(Y)$ < Max

* Only for positive control

1.3.2 Issues in interpreting and estimating variability between runs and between laboratories

1.3.2.1 Terminology

Terminology for specifying data variabilities is somewhat confusing because there is a hierarchical structure in the data from receptor binding experiments and it is not always clear to which level of the structure a particular term is referring. Even for a seemingly straightforward term such as "between-run variability" there is more than one interpretation. Consistent use of clearly-defined terminology is essential in discussion on data variability in our context.

The confusion arises from the fact that the variability observed across multiple units of observation such as runs or laboratories generally has two components. To illustrate this,

let us take a look at overall between-run variability. Under the assumption of independence between the two right-hand side terms, the following relationship generally holds:

Overall (total) between-run variability = intrinsic between-run variability + within-run variability

To understand this properly, it is illuminating to consider a special case where the true "intrinsic between-run variability" is zero. That is, all runs have a common true mean and the sole source of variation is within-run variation (data with this feature can be artificially created by performing an experiment in nine replicates then dividing the entire set into three sets of triplicate data each.) In this special case, although the true intrinsic between-run variability is zero, the observed overall between-run variability usually takes a non-zero value. The true overall between-run variability in this situation is equal to the true within-run variability.

In more general situations where intrinsic between-run variability is not zero, the observed overall between-run variability, assessed from variability across three run-specific means, most of the time would not be greater than the observed within-run variability. In principle, we can estimate the intrinsic between-run variability by estimating the overall between-run variability from the observed data first and then subtracting from it the contribution of within-run variability. It is possible that the observed between-run variability may even be smaller than the level expected from the observed within-run variation. In that case, our estimate of intrinsic between-run variability is set or truncated to zero.

Turning our attention to between-lab variability, we have:

Overall (total) between-lab variability = intrinsic between-lab variability + overall within-lab variability

The following equation relates the two equations above to each other.

Overall (total) between-run variability = Overall within-lab variability

By combining the three equations given above, we get:

Overall (total) between-lab variability = intrinsic between-lab variability + intrinsic between-run variability + overall within-run variability

That is, overall between variability at a level of the hierarchy is addition of the *intrinsic between* variability of the hierarchical level in question, *intrinsic between* variability at the lower level(s) of hierarchy, and *within* variability at the lowest level of hierarchy.

In discussion of between-run or between-lab variability, the distinction between the overall between variability and intrinsic between variability often is vague. There is a tendency for a non-statistical data analyst to interpret the observed overall between-run variability as a proxy for intrinsic between-run variability. The observed overall between-run variability and intrinsic between-run variability are generally not the same when within-run variability is large relative to intrinsic between-run variability. Thus, the contribution of the within-run variability to the overall between-run variability cannot be ignored. Similar caution applies to overall between-lab variability, intrinsic between-lab variability, and within-lab variability.

1.3.2.2 Limitations in the conventional approach to estimating variability between laboratories

One of the general goals in assay validation is to assure low data variability. An ideal assay would have low variability at all hierarchical levels of analysis. One way to assure this is to establish a low level of overall between-lab variability, which is a sufficient condition for low variability at all levels. Low overall between-lab variability implies that all of the three components (intrinsic between-lab variability, intrinsic between-run variability, overall within-run variability) are low. When this is true, a method for assessing variability that is primarily based on the overall between-lab variability is conceivable.

This approach of focusing only on overall between-lab variability, which shall be referred to as the conventional approach hereafter, has a few limitations. The first limitation is that it provides no information about where the source of variability lies. If overall between-lab variation is found to be high, we would conclude that additional efforts to improve the assay in terms of variability is necessary, but merely knowing that the overall between-lab variability is high does not provide any guidance as to how to control it. With additional information on a major source(s) of the variability, it would be possible to target efforts towards reducing variability in the source(s). Knowing the relative contributions of the three aforementioned sources of variability is useful for such targeting.

There is another shortcoming of the conventional approach. If we use the observed overall between-lab variability alone, ignoring the observed within-lab variability when estimating the overall between-lab variability, we tend to underestimate the true underlying level of overall between-lab variability in certain situations (to be discussed in detail in "2.8.2.2 Underestimation of between-unit variability in the conventional method").

1.4 Proposed solutions

1.4.1 Estimate between-lab variability via the DerSimonian/Laird model

A relatively simple statistical method is useful in overcoming the limitations described above. The DerSimonian and Laird random effects model (DL model) (DerSimonian and

Laird, 1986) allows estimation of intrinsic between-unit variability, which then can be combined with conventional estimates of within-unit variability to produce a good estimate of overall between-unit variability.

Let's take an example of summarizing $\log IC_{50}$ across runs (and then across labs) to illustrate the use of the DL model. The DL method can be used to summarize estimates of the three other parameters of interest, i.e., top plateau level, bottom plateau level, and Hill slope. Suppose we have the following data from a single lab. SE($\log IC_{50}$) is standard error of "$\log IC_{50}$ for each run" or "run-specific $\log IC_{50}$".

Run	$\log IC_{50}$	SE($\log IC_{50}$)
1	-9.02	0.40
2	-9.41	0.33
3	-8.50	0.21

Fitting the DL model to these data generates an overall summary $\log IC_{50}$ of -8.93 with SE of 0.30. Intrinsic between-run variability expressed as standard deviation is estimated to be 0.42. This estimate is not available through the conventional approach, which would compute standard deviation (standard error) for three $\log IC_{50}$ observations as an estimate of overall between-run variability using a regular formula, ignoring SE($\log IC_{50}$) of each run. Exactly how the DL model computes these is to be discussed in Chapter 2.

Let's say the above results were reported by lab 1. Suppose data from two other laboratories (lab 2 and lab 3) also are summarized in a similar manner, and the following is obtained. (SE($\log IC_{50}$) denotes standard error of "$\log IC_{50}$ for each lab" or "lab-specific $\log IC_{50}$"):

Lab	$\log IC_{50}$	SE($\log IC_{50}$)
1	-8.93	0.30
2	-9.20	0.15
3	-8.71	0.14

The same method that was used to summarize results across runs can be applied to these results to summarize results across labs, giving an overall summary $\log IC_{50}$ of -8.95 with SE of 0.18. Intrinsic between-lab variability expressed as standard deviation is estimated to be 0.24. This estimate of intrinsic between-lab variability again is not available through the conventional approach.[8]

[8] In the conventional approach, estimates of overall between-run variability and overall between-lab variability are computed without using run- or lab-specific SE($\log IC_{50}$). In order to grasp a potential problem with this disregard, consider another data set for three labs similar to but different from the one shown above. Suppose this second data set has the same $\log IC_{50}$ but its SE($\log IC_{50}$) values are the above values multiplied by a very large number. Note that the multiplication factor could be set large enough to make the estimate of intrinsic between-lab variability to be truncated to zero. If we further increase the multiplication factor, the estimate of overall between-lab variability could be made arbitrarily large, greater than the level estimated by the conventional method. Let's say the second data set has such a huge within-lab variability. From these two data sets, the conventional approach would compute the same estimate of overall between-lab variability no matter how large the within-lab variability is. Intuitively, the overall between-lab variability in the second data set should be greater than that in the first data set because of much greater within-lab variability in the second data set. As explained in detail in "2.8.2.2

The SE is for the mean over three labs; therefore, the standard deviation (SD) of $\log IC_{50}$ reported by any given lab drawn from the universe of laboratories from which these three labs are selected is 0.18 * sqrt(3). In general, if the SE for the overall mean is calculated across *k* labs,

$$\text{SD(lab-specific logIC50)} = \text{SE(pooled mean logIC50 from } k \text{ labs)} * \text{sqrt}(k) \quad \text{.....} \quad 1.3^{9}$$

The measure of variability we use here is SD, which describes the distribution of $\log IC_{50}$s reported by individual labs, not the distribution of the pooled mean of the $\log IC_{50}$s from multiple labs.

1.4.2 Deriving accuracy performance criteria

Let's say we have data from several laboratories that we deem acceptable and calculated SD($\log IC_{50}$). Using this SD, which is a measure of between-lab variability, along with the overall mean $\log IC_{50}$, we can describe a distribution of $\log IC_{50}$s reported by a lab drawn from a universe of laboratories from which these several labs were drawn. From this distribution, we can expect that the $\log IC_{50}$ from a lab from this universe would fall in the following range 80% of the time.[10]

$$(\text{Overall mean } \log IC_{50}) \pm (t_{0\,90} * \sqrt{1+1/k} * \text{SD(lab-specific } \log IC_{50})) \quad \text{.....} \quad 1.4^{11}$$

Underestimation of between-unit variability in the conventional method", the conventional approach-based estimate of the overall between-lab variability is likely to underestimate the true underlying level of overall between-lab variability when within-lab variability is relatively large compared to the between-lab variability, like in the second data set.

[9] Dr. Feder pointed commented as follows (Feder, 2007b). "This is approximation. It is exact when the within-lab component of variability is constant across labs. Otherwise, strictly speaking a standard deviation is not really defined." It would be worthwhile to see if the exact method (one way heterogeneous ANOVA) proposed by Dr. Feder (2007c), which does not assume within-lab variability being constant across labs, would generate much different results. If the difference is small, it would be justified to use the approximation. Also, although heterogeneity of within-lab variability is noticeable in our current data for $\log IC_{50}$ whether the statistical evidence for the heterogeneity is strong or not has not been assessed to my knowledge. The argument for using the exact method would be strengthened by empirically demonstrating the violation of the constant variability assumption using real data. It is suspect that the power to detect heterogeneity may be rather limited given the size of available data, i.e., data are available only up to four labs.

As a related topic, to my knowledge the assumption of constant within-run variability across runs within a lab has not investigated either. Since there are 3 to 12 runs per lab, the power for detecting across-run heterogeneity would be better than that for detecting across-lab heterogeneity. If the assumption is found to be violated, the use of one way heterogeneous ANOVA would also be necessary when summarizing results across runs within a lab.

[10] Different levels of probability coverage for different parameters. Our original choice of 80% for $\log IC_{50}$, rather than 95% that was used for top, bottom, and Hill slope parameters, was the result of the EDSP's initial desire to tightly control accuracy of $\log IC_{50}$. When the probability of 90% or 95% was used for deriving the acceptable range for $\log IC_{50}$, the width of range appeared to be too wide and unsatisfactory. In contract, the range based on the 80% probability coverage appeared satisfactory.

[11] The formula 1.4, which is a prediction interval, was a corrected version proposed by Dr. Feder. The following formula that appeared in previous drafts,

We can compare a particular $\log IC_{50}$ reported from a new lab to this range to decide whether we would deem it as acceptable or not. Either of the following actions may be taken based on this comparison:

- Accept the lab if the lab's $\log IC_{50}$ falls in the range
- Reject the lab if the lab's $\log IC_{50}$ falls outside of the range

This would ensure that we would be accepting a lab, with a probability of 80%, if it belongs to the set of all labs which are similar in competence to the labs which generated the data from which the $\log IC_{50}$ and SD performance criteria were derived.

With some necessary modifications to how to estimate SD, this general approach may also be applied to logRBA. [12]

In order to set performance criteria for estimates of the three other parameters of interest, e.g., top plateau level, a slightly modified version of the above-described method can be used. As mentioned earlier, the estimate of these parameters for each run should be within a relatively narrow range (for typically behaving one-site competitive binders such as the reference chemicals), and imposing performance criteria on an estimate from a single run, rather than an estimate based on multiple runs from a lab, is desired. For example, a relevant measure of spread for deriving performance criteria for an estimate of parameter T (top plateau) would be

$$\text{SD(run-specific T)} = \text{SE(pooled mean T from 3 runs)} * \text{sqrt(3)} \quad \text{.....1.5}$$

Using this measure of spread, a performance criterion in the form of lower and upper 95% limits[13] is computed as

$$(\text{Overall mean T}) \pm (t_{0.975} * \sqrt{1+1/k} * \text{SD(run-specific T)}) \quad \text{.....1.6}$$

$$(\text{Overall mean } \log IC_{50}) \pm (Z_{0.90} * \text{SD(lab-specific } \log IC_{50}))$$

is incorrect. This is the most important correction proposed by Dr. Feder and the reader of this document should avoid the use of the incorrect formula. The use of $t_{0.975}$ instead of $Z_{0.975}$ is suggested because the SD is unknown and estimated from the data. The degree of freedom based on Satterthwaite-type approximation is after Hartung and Makambi (2001) according to Feder (2007a, 2007b, 2007c). The inclusion of 1/k in the square root term is to account for uncertainty around the estimated mean.

[11] Dr. Feder stated that $\log(SE(\log IC_{50}))$ "is more nearly Normally distributed [than $SE(\log IC_{50})$], and so the asymptotic method will work better in small sample sizes."

[12] To be described in "2.5.2 logRBA".

[13] Some rationale for choosing the probability coverage among alternatives, e.g., 95% vs. 80%, is explained in "3.5.5 logRBA". A notable difference between the above-mentioned example of $\log IC_{50}$ and the example of T (top plateau parameter) is that in the former the lower and upper limits were derived for a pooled mean of $\log IC_{50}$s from 3 runs while in the latter the lower and upper limits were derived for a top plateau estimate from a single run.

1.4.3 Deriving precision performance criteria

As mentioned earlier, we also need precision criteria for $\log IC_{50}$, logRBA, and within-run variability of individual y measurements.

1.4.3.1 Performance criteria for variability of $\log IC_{50}$ and logRBA estimates

A measure of within-run variability for $\log IC_{50}$ is $SE(\log IC_{50})$ reported for a single run. Since $SE(\log IC_{50}) > 0$, it makes sense to work on $\log(SE(\log IC_{50}))$,[14] which we shall call λ. We can impose an upper limit for λ to ensure acceptable precision for $\log IC_{50}$. Since small λ is a good feature, there is no need to impose a lower limit for λ.

Theoretically λ can be summarized across runs and across labs in a manner similar to how we establish limits for four parameters in the Hill equation model. An problem here is that $SE(\lambda)$ is not readily available in output from a nonlinear regression. In order to overcome this, a Jack-Knife method of estimating variance may be used.[15] A detailed discussion of this method will be given in "2.2.4 Jack-Knife variance estimation".[16]

Once SE(lab-specific λ) is obtained, a performance criterion for λ in the form of an upper limit for 90% prediction interval can be set as

$$(\text{Overall mean } \lambda) + (t_{0\,90} * \sqrt{1+1/k} * \text{SD(lab-specific } \lambda)) \quad \text{.....1.7}$$

Procedures for deriving an upper limit for log(SE(logRBA)) are similar.

1.4.3.2 Performance criteria for within-replicate-set variability of y [17]

Another measure of precision on which we would like to impose an upper limit is within-replicate-set variability of individual y measurements. This within-replicate-set variability can be represented by $\log(SD_{\text{within-replicate-set}}(Y))$, which we call ν (nu).

---

[14] Dr. Feder commented $\log(SE(\log IC_{50}))$ "is more nearly Normally distributed [than $SE(\log IC_{50})$], and so the asymptotic method will work better in small sample sizes."

[15] Dr. Feder proposes the use of delta method in place of Jack-Knife method (Feder, 2007b, page 27). I agree with his preference for delta method based on its ease of computation. This point applies to all occurrence of the proposed uses of Jack-Knife throughout the document. Although delta method as used in this context is considered as a simple procedure by DCC, it involves the use of not widely-known Satterthwaite approximated degrees of freedom after Hartung and Makambi (2001). If the Hartung and Makambi method is to be used by DCC on a regular basis, EDSP would benefit from having easy-to-follow instructions as to how it is performed.

[16] The standard error of a parameter estimate is the square root of the variance of the parameter estimate.

[17] Please note that in the revision of performance criteria numbers mentioned in footnote 1 the approaches described in Sections 1.4.3.2 and 1.4.3.3 was to be largely abandoned and an alternative method similar to the one used for the four Hill equation parameters was to replace those approaches. Dr. Feder thought that the justification for the methods described in Sections 1.4.3.2 and 1.4.3.3 was unclear. These methods were developed based primarily on EDSP's desire in the past to control precision of $\log IC_{50}$ estimates. Now EDSP has lost such a desire, it makes sense to use a method comparable to the one chosen for other parameters.

An approach somewhat different from the approaches taken for other parameters was taken to derive the upper limit for ν. In short, the limit can be set by translating the upper limit for λ (i.e., log(SE(logIC$_{50}$))) to that for ν using an empirical relationship between ν and λ.

Through analysis of historical data for the RUC ER binding assay it was noted that ν and λ are positively correlated. Since the parameter estimate of ultimate interest is logIC$_{50}$, our main target for precision control is log(SE(logIC$_{50}$)), i.e., λ. To the extent that ν and λ are positively correlated, keeping ν low ensures low λ as well.

As discussed earlier, a method for deriving an upper limit for λ has been developed. The positive correlation between ν and λ allows us to translate the upper limit for λ into that for ν.

1.4.3.3 Performance criteria for overall and between-replicate-set variabilities of y given x

We also consider setting upper limits for total variability and intrinsic between-replicate-set variability of individual measurements of binding at a given concentration. The relationship between these two variability measures and within-triplicate-set variability, which was discussed in the previous section, is similar to that described for SE(logIC$_{50}$), that is,

Overall (total) variability = intrinsic between-replicate-set variability
+ within-replicate-set variability

Intuitively, the total variability of binding measurements at a given concentration would be most closely associated with the standard error of the lab-specific logIC$_{50}$. This is because the standard error of the lab-specific logIC$_{50}$ is a function of mean square errors from the nonlinear regression fitted to the data. The mean square errors would equal the total variability of y given x as long as the condition that the mean of y at a given x on average equals the level expected from the underlying curve.[18] This forms a basis for setting up an upper limit for the total variability in binding at a given concentration.

It was noted while analyzing the historical data that the intrinsic between-run variability in logIC$_{50}$ contributed more to the within-lab variability of the logIC$_{50}$ than did its other component, within-run variability. Intuitively the between-replicate-set variability be more closely correlated with the intrinsic between-run variability in logIC$_{50}$ than within-replicate-set variability. Therefore, setting an upper limit for the between-replicate set variability is also an option.

[18] This condition may not hold if the expectation of Y given x systematically differs from the expected level in the underlying curve. Such systematic difference is possible, but there is no strong reason to suspect that such differences exist. It would be of interest to investigate this in the context of regression diagnostics.

A counter-argument for setting these upper limits is that SE(lab-specific $\log IC_{50}$) can be estimated with the same data that are used in estimating between-replicate-set variability in binding at a given concentration. If keeping SE(lab-specific $\log IC_{50}$) low is one of our ultimate goals, directly controlling it rather than through controlling total or intrinsic between-replicate-set variability in binding at a given concentration may be more straightforward and justifiable.

Nonetheless, the capacity to estimate within-run and between-run, as well as the total variability in binding at a given concentration, may be useful when it is difficult to locate laboratories that have sufficiently small standard errors on the $\log IC_{50}$. Knowledge of an individual laboratory's within-replicate-set and between-replicate-set variability in binding measurements at a given concentration allows us to modify the protocol for a substandard lab, such that it would generate sufficiently precise data (i.e., require additional replications if its within-replicate-set variability is exceedingly high, or require additional runs if its between-replicate-set variability is exceedingly high).[19]

[19] Dr. Feder commented on this paragraph as follows. "I disagree. A lab cannot make up for producing imprecise or inaccurate results by producing more of them! Better to find out why the lab is more variable. Standard assay should require a specified number of replications." I am not proposing that we should routinely allow labs to increase number of replications/runs as a kind of "loophole" for poorly-performing labs. This is proposed as a last resort when an agency is desperate to secure many enough acceptable labs. This situation might arise if data variability remains to be too high even after exhaustive attempts to improve an assay protocol. In such a case, it might be necessary to increase the specified number of replications/runs for any lab. I agree it would be better to investigate the reason for a exceeding high variability for a given lab, but with limited resources an agency may not be able to do so.

2 Statistical Methods

2.1 Basic assumptions

2.1.1 Input data specification

Input data for the dependent variables should be standardized and expressed as "% binding of the reference ligand to the receptor". Concentration of a standard/test chemical is expressed as log(concentration). Generally speaking, the use of concentrations and IC_{50}s on absolute scales shall be avoided, primarily because the modeling equation of choice includes log(concentration), not concentration itself (see modified Hill equation, equation 1.2).[20] By fitting the model to the data using nonlinear regression, we obtain an estimate of the $\log IC_{50}$, not the IC_{50} itself, along with its standard error.[21]

The chemical concentration levels and the number of replicates at each concentration must be specified for the chemicals (viz., reference standard and positive control) for which performance criteria are to be set. If data were collected at a greater number of concentration levels than specified in the protocol, the standard error for each of the parameters of interest on average would be smaller. The same is true if a greater number of replications were used at each concentration. Also, some concentration levels are more informative in estimating certain parameters. For instance, data collected at concentration levels near the true $\log IC_{50}$ are particularly informative for $\log IC_{50}$ estimation, and data collected at very low concentration levels are generally more informative for estimating top plateau levels. In general, data to be used for describing data variability and setting performance criteria should conform to protocol specifications regarding concentration levels and number of within-run replications at each concentration.

2.1.2 Method for model fitting

2.1.2.1 Model equation

Conventionally the following equation, called the 4 parameter Hill equation or 4 parameter logistic equation, is used to describe single-site competitive binding data.

---

[20] Dr. Feder comments as follows. "No! I agree with the statement but not the rationale. Distribution of $\log(IC_{50})$ is more nearly normal than that of IC_{50} so asymptotic distribution theory works better." I do not think my version of rationale is incorrect. I argue that the reason why distribution of $\log(IC_{50})$ is more nearly Normal is because the Hill equation has log(**concentration**) not **concentration** as a right hand-side variable. The log(concentration) is used in part because its use would make the distribution of residuals more nearly Normal. That is, the two versions of rationale complement, rather than contradict, with each other.

[21] There is a tendency for experimentalists to prefer expressing concentration on absolute scales, including the use of IC_{50} rather than $\log IC_{50}$. In order to summarize IC_{50} estimates across labs, we would need $SE(IC_{50})$. Note that is it incorrect to compute $SE(IC_{50})$ as $10^{SE(\log IC_{50})}$ We can more correctly obtain $SE(IC_{50})$ using a delta method, but the $SE(IC_{50})$ obtained thereby is not ideal, since the sampling distribution of the IC_{50} would not be symmetrical. It would be acceptable to perform all the analyses on the log scale and translate the final results back into IC_{50}. Dr. Feder recommends that we calculate "boundary for $\log IC_{50}$ and then exponentiate them" and asserts that "[t]his is the standard approach. This approach is exactly what I meant in the previous sentence.

$$Y = B + \frac{T - B}{1 + 10^{(\mu' - X)*H}} \quad \text{.....1.1}$$

where μ' = logEC$_{50}$, H = Hill slope, T= top plateau level, B = bottom plateau. EC$_{50}$ is the "effective concentration, 50%"—that is, the concentration at which binding = (B+T)/2, and is represented graphically as the x-coordinate of the vertical mid-point of the curve. This parameter is of limited use for chemicals for which the bottom plateau is not 0 % and the top plateau is not 100%, because the % binding corresponding to the EC$_{50}$ changes according to the locations of the bottom and top plateau.

In order to compare binding affinities of different chemicals, we would like a summary measure that allows one to represent the concentration associated with a specific, standardized level of binding. For a level of 50%, this concentration is called the IC$_{50}$ ("inhibitory concentration, 50%") and represents the concentration at which 50% of a reference ligand is displaced from the receptor by the competitor. It can be estimated by fitting the binding data to the following equation.

$$Y = B + \frac{T - B}{1 + 10^{[(\mu - X)*H + \log(\frac{T-B}{50-B} - 1)]}} \quad \text{.....1.2}$$

where μ = logIC$_{50}$.

The difference between logEC$_{50}$ and logIC$_{50}$, and the reason that logIC$_{50}$ is the preferred quantity to use for comparing binding affinities across chemicals, are illustrated in Figure 2.1. The four solid curves shown all have the same logEC$_{50}$ at -7.5 and would be regarded as having the same binding affinity if we use logEC$_{50}$ as a measure of affinity. They, however, differ in terms of logIC$_{50}$. No logIC$_{50}$ exists for the dashed curve (see the next paragraph). The three solid curves have different logIC$_{50}$s with the chemical having the lowest bottom plateau also having the lowest concentration at which 50% of the reference ligand is displaced from the receptor (that is, the smallest logIC$_{50}$).

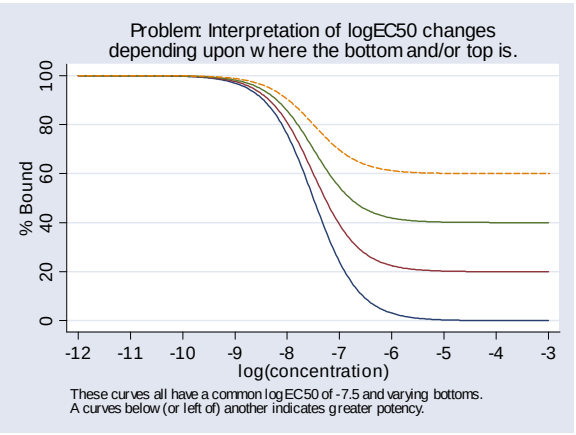

Figure 2.1 Limitation of logEC$_{50}$

It should be noted, though, that the logIC$_{50}$ is not perfect since we cannot estimate logIC$_{50}$ for a chemical whose curve does not cross the 50% level of reference-ligand binding (in the figure, the upper most, dashed curve), and we will not be able to compare its logIC$_{50}$ to those for the other curves in figure 2.1. In other words, the logIC$_{50}$-based

model cannot be fit to the data for a chemical of this sort. A $logEC_{50}$-based model still could be fit to the data of this kind, allowing us to document bottom, top, and slope of the curve. The two approaches, one based on $logIC_{50}$ and the other on $logEC_{50}$, are complementary and have their own merits and demerits.

The problem associated with ranking chemicals using $logEC_{50}$ is illustrated again in Figure 2.2 using two underlying curves with different Hill slope values.

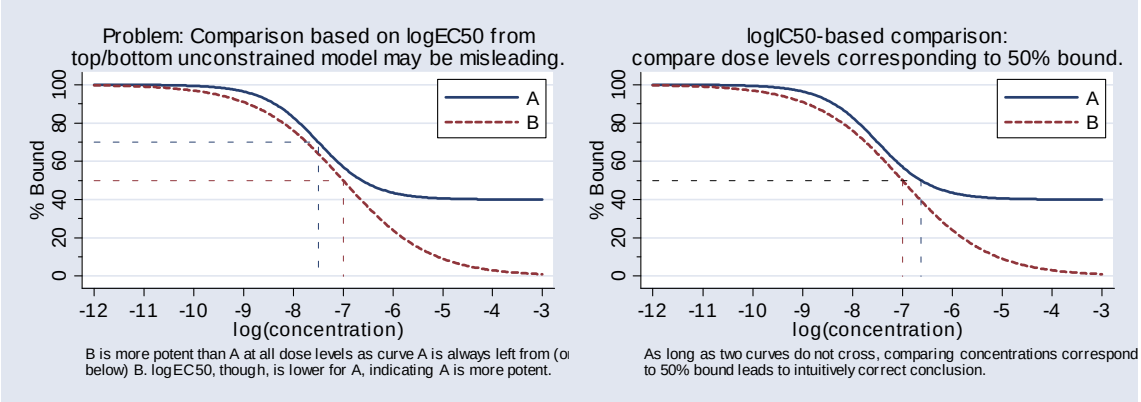

Figure 2.2 Advantage of $logIC_{50}$ over $logEC_{50}$

Curve A is above (or right) of the curve B at any x, indicating the former is less potent than the other. The $logIC_{50}$ for A is higher than that for B, properly reflecting this relative relationship between the two curves. The $logEC_{50}$ for A, however, is lower than that for B and if it were being used as the indicator of relative binding affinity we would improperly conclude that the chemical represented by curve B is a stronger binder than the chemical represented by curve A.

2.1.2.2 Parameter estimation by nonlinear regression

In fitting either model to the data and estimating the four parameters, a nonlinear regression method is used. This procedure usually is available in regular commercial software. Using the $logIC_{50}$-based equation, the statistical model for the nonlinear regression is expressed as follows,

$$Y = B + \frac{T - B}{1 + 10^{[(\mu - X)*H + \log(\frac{T-B}{50-B} - 1)]}} + \varepsilon \qquad \text{.....2.1}$$

where ε is an error term. Typically, parameter estimation is performed via the least squares procedure, i.e., by finding the values of the four parameters that minimize the residual sum of squares, assuming ε is independently and identically normally distributed.[22]

[22] Some software provide an option for log-normal error or other error structure, but as explained in "2.1.2.5 Homoskedastic error structure" below we chose to assume Normal distribution.

The least squares method may not be the best estimation procedure for receptor binding data because it treats all data points equally even though the data may contain outliers (or exceptional data points), which intuitively should be somehow downplayed. By giving smaller weight to outliers we may be able to improve the accuracy and precision of the estimate. This may be achieved by using a so-called "robust" regression method such as the iterative re-weighted least square method. Investigation of whether the use of robust regression would considerably improve estimation of parameters from datasets likely to be encountered in screening chemicals for receptor binding is beyond the scope of this report.

It may be argued that fitting the $\log EC_{50}$-based model rather than the $\log IC_{50}$-based model would be preferable since that model could be fit irrespective of the bottom plateau level. This position is aided by the fact that the $\log IC_{50}$ and its standard error may be derived from the $\log EC_{50}$ and the other three parameters as

$$\log \mathrm{IC50} = \hat{\mu} = \hat{\mu}' - \frac{\log\left((\hat{T} - 50)/(50 - \hat{B})\right)}{\hat{H}} \quad \text{.....2.2}$$

where $\hat{\mu}'$ = $\log EC_{50}$. SE($\hat{\mu}$) may be derived by the delta method. It has been confirmed that the $\log IC_{50}$ results from $\log IC_{50}$-based regression and those from $\log EC_{50}$-based regression followed by the delta method agree completely at least in Stata, a statistical package. The same is likely to be the case for other statistical packages.

The reason we did not choose this $\log EC_{50}$-model-based $\log IC_{50}$ estimation as the default approach is a practical one. Although the delta method is usually available in regular commercial statistical software packages such as SAS and Stata, these packages may not always be accessible to the laboratories that perform the receptor binding assay. For instance, a curve-fitting package called GraphPad Prism (version 4) that is widely used for analysis of receptor binding data does not have this generic delta method.

Although there is a practical disadvantage to the approach based on the $\log EC_{50}$-based regression equation, its use should be kept in mind.

2.1.2.3 Constraints for parameters

Commercial software for model-fitting sometimes includes an option of "fixing" certain parameters in nonlinear regression by way of a built-in, simplified form of a general equation. For instance, the bottom, top, and slope parameters could be fixed to 0, 100, and -1, respectively, giving rise to the following equation.

$$Y = \frac{100}{1 + 10^{X - \mu'}} \quad \text{.....2.3}$$

This kind of option often is available also under an option of "parameter constraints", which allows us to force a parameter to be a constant or to fall in a range, e.g., bottom > -10%. In general, constraints like this should be avoided [23] Their use gives a false sense of security regarding the precision of parameters for which no constraint was imposed. In certain situations, constraints imposed on parameters may bias estimates of other parameters. [24]

2.1.2.4 Error structure assumption

As mentioned earlier, independently and identically distributed (iid) Normal error structure is assumed in the regular form of nonlinear regression we apply to receptor binding data. Other forms of error structure such as log-Normal and Poisson are possible.

For a given error structure the best weighting scheme is implied. No weighting or equal weighting is the best weighting scheme for iid Normal error. For log-Normal and Poisson distributions of error, weighting by $1/Y^2$ and $1/Y$, respectively, is optimal.

EDSP has found it justifiable to assume that the error structure of receptor binding data is Normal rather than log-Normal or Poisson, and to use equal weighting. The basis for this position is two-fold. The first, a practical reason, is that the alternative error structures, where the error is assumed to be proportionate to a function of the response variable, could not be handled properly in certain widely used commercial software such as GraphPad Prism.[25] The second reason, an empirical one, is that there is evidence that the error structure is better described as Normal rather than Poisson or log-Normal.

Practical reason: The alternative error structure calls for a weighted nonlinear regression. Weights are supposed to be positive, but the actual weights may be calculated as a negative number when an observed response is negative. This happens when a Poisson error structure is specified and there are negative y values. When faced with negative weights, software may stop, issuing an error message.

---

[23] Dr. Feder comments as follows. "We can do goodness-of-fit test to see if the constraints are appropriate." If this is a suggestion that goodness-of-fit test be performed routinely using some selected (set of) constraints such as top = 100% or slope = -1, then I would disagree mainly because of the reason stated in the following sentence. Routinely applying different constraints and testing which constraints are appropriate also seem impractical.

[24] Dr/ Feder suggests that the phrase, "(only if the constraints are not valid)", be inserted here.

[25] Dr. Feder states "This is not a good reason… There are workarounds in Prism…" I believe the "workaround" referred here is the method of adding a (arbitrary) small positive constant (say, c) to all % binding values and specify weighting scheme of choice in nonlinear regression for fitting a Hill equation. This would require estimating $\log IC_{50+c}$ in place of $\log IC_{50}$. The procedure for choosing an appropriate value of the constant c is not self-evident. The constant needs to be greater than absolute(the smallest observed % binding value") when the smallest observed value is negative. It does not seem justifiable to use a pre-set constant for all experiments for different labs as error structures would very across labs (and potentially between runs, within lab). These would mean the constant needs to be chosen for each run (or lab). There would be certain arbitrariness in the process of choosing the constant. The approach of using non-weighted procedure also would be deemed arbitrary given uncertainly regarding the true error structure. Given these considerations, it seems that the use of "adding a positive constant" procedure seems excessively cumbersome while it is not certain how much reduction in arbitrariness its use can achieve.

Empirical reason: In order to assess error structure, we would fit an appropriate model, obtain residuals, and examine the relationship between the squared error at a given concentration and E(Y) at the concentration. (See the next section for how this is done.) When this relationship was examined in the historical RUC data, there was little evidence that the error structure was log-Normal. The observed structure was something in-between Poisson and Normal, slightly closer to Normal. As such, SD($y-\hat{y}$) is assumed to be constant across varying y.[26]

2.1.2.5 Homoskedastic error structure

The iid Normal errors represent a kind of homoskedastic error structure. Deviations from this, e.g., Poisson or log-Normal error distribution, where the magnitude of errors varies as a function of the dependent and/or independent variable, represent heteroskedasticity. Our assessment of homoskedasticity focused on whether the errors change monotonically with the dependent variable. It was performed by taking the following steps.

1. Fit the $\log IC_{50}$-based Hill equation model to the data from a run using a non-weighted nonlinear regression
2. Obtain the predicted value ($\hat{y}$) at each x of a given run
3. Obtain the residuals (= $y-\hat{y}$)
4. Obtain mean squared error at each x of a given run
5. Repeat this for all runs for all laboratories and chemicals under consideration
6. If there are q concentration levels for each run and there are a total of r runs, we have $q*r$ pairs of mean squared error and $\hat{y}^2$
7. Take the logarithm of mean squared error (MSE)
8. Take the logarithm of $\hat{y}^2$
9. Investigate relationship between log(MSE) and log($\hat{y}^2$). Specifically, regress log(MSE) and log($\hat{y}^2$) and estimate the regression coefficient k in $\log(MSE) = \beta_0 + k*\log(\hat{y}^2) + \varepsilon$

The regression coefficient k is interpreted as follows:

- k = 0 indicates the error structure is homoskedastic (error level does not change linearly with y).
- k = 0.5 indicates the error structure is Poisson-type (variance(residual) increases proportionately with $\hat{y}$).
- k =1 indicates the error structure is log-Normal-type (SD(residual) increases proportionately with $\hat{y}$).

To see this, the following rearrangement of expression for SD(residual) is informative.

[26] Dr. Feder states that the reasoning in this paragraph is "OK" for ER binding assay. He then recommends that the error structure be investigated for different assay. I agree with this as long as EDSP's resources allows such investigation. Another approach may be to perform a simulation study, and investigate whether incorrect assumption on error structure would lead to substantially different parameter estimates and their standard errors.

$$SD(residual) = \mathrm{c} \cdot \hat{\mathrm{y}}^{\mathrm{k}} \quad \text{.....2.4}$$

Squaring both sides of the above yields

$$Var(residual) = MSE = \mathrm{c}^2 \cdot \hat{\mathrm{y}}^{2\mathrm{k}} \quad \text{.....2.5}$$

Taking logarithms of both sides the above then yields

$$\log(MSE) = 2 \cdot \log(\mathrm{c}) + k \cdot \log(\hat{\mathrm{y}}^2) \quad \text{.....2.6}$$

A summary of the analysis of the noise levels for historical RUC ER binding assay data is presented in the section "2.8.1 Constant noise assumption".

2.1.3 Statistical software

There are several potential choices for statistical software to be used when fitting the 4-parameter Hill equation model to the data. They include: GraphPad Prism, which is developed primarily for laboratory data and is widely used by academic and commercial laboratories; SAS, a versatile package widely used in pharmaceutical industry and academia, which is relatively inaccessible because of high cost and; Stata, another multi-purpose package popular among public health researchers and econometrists, which probably is not as widely used as SAS but is more accessible than SAS in terms of cost. As long as nonlinear least square regression capacity is available, any other package would be usable for fitting the 4-parameter Hill equation model.

For the kind of data to be encountered in validation, i.e., data for standard and positive control chemicals, any of the three software packages listed above would produce equivalent results. For certain atypical, difficult-to-fit data that have been encountered (and might be encountered in large-scale testing), Stata and SAS performed better (and are expected to perform better) than GraphPad Prism in that estimation convergence was achieved in Stata and SAS for such data while GraphPad Prism issued an error message without completing the estimation. GraphPad Prism and Stata do not allow the full execution of the alternative model-fitting and data-summarizing method discussed in the section "2.8.2.1 Comparing methods for summarizing data from multiple runs".

2.2 Method for between-run and between-lab summary of parameters

2.2.1 DerSimonian-Laird random effects model

In dealing with summary statistics for receptor binding experiments, we encounter the need to summarize parameter estimates from independent experimental units (either run or lab). Generally, the i^{th} unit has a parameter estimate $\hat{\theta}_i$ and its standard error $SE(\hat{\theta}_i)$. For instance, for i = 1, 2, 3, we have a data set of

estimate Standard error of estimate

$\hat{\theta}_1$ $SE(\hat{\theta}_1)$

$\hat{\theta}_2$ $SE(\hat{\theta}_2)$

$\hat{\theta}_3$ $SE(\hat{\theta}_3)$

Our goal is to summarize these across units into a pooled estimate and its standard error. In doing so, we assume the following.

$$\hat{\theta}_i \sim N(\theta_i, v_i) \text{ and } \theta_i \sim N(\theta_R, \tau^2) \quad \text{.....2.7}$$

That is, two distinct sources of variation are assumed to exist. First, each of the parameter estimates from a unit has a unit-specific true value and a within-unit variation. In addition, the true parameter follows a Normal distribution with a common mean (θ_R) and common between-unit variation (τ^2). Commonly used estimators of τ^2 and θ_R are due to DerSimonian and Laird (1986). The model they proposed is called the DerSimonian-Laird (DL) random effects model. τ^2 is estimated by the method of moments as follows:

$$\hat{\tau}^2 = \max\left(0, \frac{\text{Q-(k-1)}}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}\right) \quad \text{.....2.8}$$

where

$$\text{Q} = \sum_{i=1}^{k} w_i (\hat{\theta}_i - \hat{\theta}_{\text{F}})^2 \quad \text{.....2.9}$$

$$w_i = 1/\hat{v}_i \quad \text{.....2.10}$$

$$\hat{\theta}_{\text{F}} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} \quad \text{.....2.11}$$

and $\hat{v}_i$ is the estimated variance of $\hat{\theta}_i$, i.e.,

$$\hat{v}_i = SE(\hat{\theta}_i)^2 \quad \text{.....2.12}$$

Subscript i indexes units, and k is the total number of the units. Using this estimate of τ^2, the revised weight w_i^*, the estimate of the random effects overall mean θ_{R} and its standard error are calculated as follows.

$$w_i^* = \frac{1}{\hat{v}_i + \hat{\tau}^2} \quad \text{.....2.13}$$

$$\hat{\theta}_{\mathrm{R}} = \frac{\sum_{i=1}^{k} w_i^* \hat{\theta}_i}{\sum_{i=1}^{k} w_i^*} \quad \text{.....2.14}$$

$$\mathrm{var}(\hat{\theta}_{\mathrm{R}}) = \frac{1}{\sum_{i=1}^{k} w_i^*} \quad \text{.....2.15}$$

Typically, the standard error of $\hat{\theta}_{\mathrm{R}}$, rather than its variance, is reported by a statistical package.

$$SE(\hat{\theta}_{\mathrm{R}}) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}} \quad \text{.....2.16}$$

The Q statistic has a chi-square distribution of *k*-1 degrees of freedom, and is used to test the null hypothesis of $\tau = 0$, i.e., to test heterogeneity of $\hat{\theta}_i$. In the DL random effects model's application in meta-analysis, it has been noted that using the typical α level of 0.05, the test of heterogeneity is underpowered in detecting non-zero τ (Takkouche et al., 1999). It has been recommended that "because these tests often falsely fail to detect true heterogeneity, it may be advisable to use random effects models routinely" (the National Research Council, 1992). An alternative level of α = 0.1~0.15 may also be used for this test to get around the underpowered problem, but since our number of units would be small a formal test in general of limited use. Our primary use of the Q statistic is for the point estimation of parameters of our interest, not for the formal tests.[27]

The DL random effects model shall be used when summarizing results of multiple runs within a lab, yielding summary statistics for each laboratory (i.e., a lab-specific summary estimate of the parameter of interest, its standard error, and intrinsic between-run variation). Lab-specific summary estimates and their standard errors could be further summarized across labs, again using the DL random effects model.

There is a likelihood-based, rather than a moment-based, method similar to the DL method, available in SAS. In a simulation study (Feder and Ma, 2005) this maximum-likelihood random effects model was found to outperform the DL model by providing more accurately estimated standard errors of parameter estimates. However, this method does not appear to be widely available. Because of its SAS-specific availability, we did not consider its use.

[27] In addition to the overall test for τ = 0, the Q statistic may be used to test whether a particular unit(s) are "different" from the other units in terms of their estimates of the parameter of interest. The test is based on a difference in two Q statistics computed with or without the units that are hypothesized to be different. The difference has a chi-square distribution of Δ degrees of freedom, where Δ is the number of the units that are hypothesized to be different.

Of note, $SE(\hat{\theta}_i)$ is an important ingredient for estimation of the overall mean, its standard error, and the intrinsic between-unit variability. One of the reasons for working primarily with $\log IC_{50}$, rather than IC_{50}, is that $SE(\log IC_{50})$ is much better defined than $SE(IC_{50})$. Distribution of IC_{50} estimates would be right-skewed compared to that of $\log IC_{50}$. Although $SE(IC_{50})$ may be calculated via the delta method, it is a forced summary measure of the spread, which is not symmetrical around the center of distribution. From the expression of $SE(\hat{\theta}_R)$, which is a measure of total variation of $\hat{\theta}_R$, i.e.,

$$SE(\hat{\theta}_R) = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}} \quad \text{.....2.17}$$

we can see it has contributions from two sources of variation, i.e., within-unit variation $SE(\hat{\theta}_i)^2$ and between-unit variation $\hat{\tau}^2$, although they are not combined into $SE(\hat{\theta}_R)$ by simple addition.

Of note here is that this standard error describes variation of $\hat{\theta}_R = \frac{\sum_{i=1}^{k} w_i^* \hat{\theta}_i}{\sum_{i=1}^{k} w_i^*}$ or the overall mean estimate. In order to describe how "individual $\hat{\theta}_i$ to be obtained in the future" varies due to within-unit and between-unit variabilities, we would need to convert this standard error to a standard deviation as follows. (By the strict usage of statistical terminology, the following also is a standard error since it is a standard deviation of an estimate. It is called SD in order to emphasize the fact that it is variation of an estimator for an individual unit, not of an overall mean estimator based on information from multiple units.)

$$SD(\hat{\theta}) = \sqrt{\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}} \quad \text{.....2.18}^{28}$$

When the units of interest are laboratories, *k* in this formula means that there are a total of *k* labs that took part in the study. This $SD(\hat{\theta})$ represents overall (total) between-lab variability.[29]

---

[28] Dr. Feder comments as follows, repeating the point he made on Equation 1.3 (see footnote 9). "This is an approximation. It is exact only if within-lab SE, $SD(\hat{\theta}_i)$, is constant across labs. Otherwise standard deviation is not actually defined." My response to this is included in footnote 9.

[29] Please note that $SE(\hat{\theta})$ and $SD(\hat{\theta}_i)$ look similar but denote two very distinct quantities. $SD(\hat{\theta})$ is the standard deviation of individual $\hat{\theta}$ (either a run-specific or lab-specific estimate of θ) *to be obtained in*

When the units of interest are runs in a particular lab, *k* in this formula means that there are a total of *k* runs for the lab. In this case, the $SD(\hat{\theta})$ represents overall (total) between-run variability or lab-specific overall within-lab variability.

As described later, this method can be used to summarize the log(standard error of estimate) or $\log(SE(\hat{\theta}_i))$. In order to do this, we would need $SE(\log(SE(\hat{\theta}_i)))$, which typically is not reported as standard output from the DL random effects model. Estimates of $SE(\log(SE(\hat{\theta}_i)))$ may be obtained using the Jack-Knife method of variance estimation, which is described in "2.2.4 Jack-Knife variance estimation".[30]

In the context of the model to be described in "2.3.2 ~~Random effects one-way ANOVA~~", the relative contribution of intrinsic between-unit variability to the total between-unit variability may be described using a measure called the intraclass correlation coefficient, which is defined as "between variance/(between variance + within variance) or $\sigma_\alpha{}^2/(\sigma_\varepsilon{}^2+\sigma_\alpha{}^2)$ using the notation introduced at the end of "2.3.2 ~~Random effects one-way ANOVA~~". Intraclass correlation is the proportion of the total variation that is explained by the "between" variation and is comparable to the widely used R-squared statistic. By analogy, we can compute a statistic with equivalent meaning for the DL random effects model as

$$\text{Intraclass Correlation} = \hat{\tau}^2 / SD(\hat{\theta})^2 \quad \text{.....2.19}$$

2.2.2 Correction for the number of runs performed[31]

Sometimes data are available for more than 3 runs. This will be the case, for example, if estradiol and the positive control are run concurrently with many unknown test chemicals. As long as there are more chemicals to be tested than can be accommodated in a single run, we will end up with data for more than 3 runs of estradiol and positive control in our efforts to obtain 3 runs of each unknown chemical. Intuitively, when trying to summarize

---

the future for a run (lab) that is similar to the runs (labs) that were considered in deriving $\hat{\theta}_R$. $SE(\hat{\theta}_i)$ is the estimated standard error of the $\hat{\theta}_i$ *that has been observed*.

[30] In our initial attempt to summarize $\log(SE(\hat{\theta}_i))$ in 2005, we did not use the Jack-Knife method. Instead, we used simulation-based estimates of $SE(\log(SE(\hat{\theta}_i)))$, which were available in a study by Battelle (Feder and Ma, 2005). The idea to use the Jack-Knife method was conceived later, and actual analysis using the method has not been preformed. In Chapter 3 the simulation-based estimates of $SE(\log(SE(\hat{\theta}_i)))$ are given. Because Jack-Knife-based standard errors may be obtained with much less computational effort than simulation-based standard errors, the use of the Jack-Knife method is recommended in future efforts to summarize $SE(\log(SE(\hat{\theta}_i)))$. Dr. Feder recommends the use of delta method in place of Jack-Knife.

[31] Please see "4.2 Improved correction for the number of runs performed", which is an addendum written after the most of this report was completed. An improved equation to correct for varying numbers of runs per lab is proposed.

results across labs we would like $SE(\hat{\theta}_i)$ (here, *i* denotes labs) to be comparable to each other. Also, we would like to ensure that each lab is "fairly" represented in deriving the overall mean. For instance, when a lab-specific summary estimate and its standard error for lab A are based on data from 10 runs and that for lab B is based on data from 5 runs, the conditions of "comparability" and "equal representation" are not met because: $SE(\hat{\theta}_A)$ would be 1/sqrt(2) of $SE(\hat{\theta}_B)$ even if they have the same level of within-lab, between-run variability; and because lab A is represented more than lab B by a factor of 2 in the universe of labs we are envisioning (The "comparability" issue and "equal representation" issue actually are the two faces of the same thing in this example).

A simple, though imperfect, solution to this problem is to draw a random sample of the same size (e.g., 3) of runs from each lab, and derive lab-specific summaries. This is inefficient as we end up throwing out much of the available data. We do not recommend this approach.

A better solution is to use all available data, but to modify the lab-specific summaries such that they become comparable to each other, at the same time ensuring equal representation. For the summary estimate standard error for lab *i*, this is done by applying a factor of $3/k_i$ in the standard error formula as follows.

$$SE(\hat{\theta}_{\text{R for lab i}}) = \sqrt{\frac{1}{\left(\sum_{j=1}^{k_i} \frac{1}{SE(\hat{\theta}_{ij})^2 + \hat{\tau}_i^{\,2}}\right)\frac{3}{k_i}}} \quad \text{.....2.20}^{32}$$

That is, recalculate the standard error as if we obtained the overall mean from 3 runs, instead of k_i runs as actually performed, thus forcing the hypothetical 3 runs to be representative of the k_i runs. The key number in this calculation, 3, is chosen since it is the default number of runs for unknowns and represents the minimum number of runs we will have for any chemical (the number of runs for the concurrent standard and positive control will always be 3 or more).

2.2.3 Pooled estimation of within-lab variability

Each lab has its own $SE(\hat{\theta}_i)$ (here $\hat{\theta}_i$ being the lab-specific estimate of θ), which tells us about the within-lab variability for that lab. This within-lab variability is lab-specific, but we often need to estimate the typical within-unit variability for multiple units. For instance, when we evaluate variability of $\log IC_{50}$ we would like to report three measures:

---

[32] Dr. Feder notes "This is approximation for the reasons discussed above." See footnote 9. This comments implies inappropriateness of this formula, but that negative connotation contradicts with the method that Dr. Feder used in his illustration of how to construct prediction intervals and tolerance intervals (Feder, 2007c). In that illustration, he apparently uses the Equation 2.20-like correction (i.e., the column labeled "**SE Crctd 3 Runs**" in Table 1 and seems to be comfortable to assume the constant within-run variability across runs within a lab. I think he mistook this equation for the one for across-lab summary rather than across-run summary, which this actually is.

overall between-lab variability, intrinsic between-lab variability, and within-lab variability. The DL random effects model generates estimates of overall between-lab variability and intrinsic between-lab variability, but it does not generate a pooled estimate of within-lab variability common to the multiple labs. That is, estimates of lab-specific within-lab variability from each lab, or $SE(\hat{\theta}_i)$, are generated as the DL method is used for *within*-lab summary, but the DL method as a procedure for *across*-lab summary does not generate a common estimate of within-lab variability. A method for generating overall or pooled estimates of within-lab variability is needed because an estimate of within-lab variability is required in many validation guidelines. There are two approaches for this.

One approach is to work directly with $\log(SE(\hat{\theta}_i))$ and its standard error, rather than $\hat{\theta}_i$ and $SE(\hat{\theta}_i)$, and fit the DL random effects model to $\log(SE(\hat{\theta}_i))$. Let us call this direct pooling. It makes sense to work with $\log(SE(\hat{\theta}_i))$ rather than $SE(\hat{\theta}_i)$ per se because $SE(\hat{\theta}_i)$ is always positive and its distribution would be right-skewed. To take this approach, estimates of $SE(\log(SE(\hat{\theta}_i))$ are needed. They may be obtained using the Jack-Knife method of variance estimation.[33]

The second approach is what we can call a "subtraction" method, which was actually used in this report. In this approach, we keep working with $\hat{\theta}_i$ and $SE(\hat{\theta}_i)$, obtain overall between-lab variability and intrinsic between-lab variability of $\hat{\theta}_i$, and from these two variability measures derive a pooled estimate of within-lab variability exploiting the relationship that exists for overall between-lab variability, intrinsic between-lab variability, and within-lab variability, which can be informally described as

Overall (total) between-lab variability = intrinsic between-lab variability + overall within-lab variability

or more formally

$$\text{var}(\hat{\theta}_i) = \hat{\tau}^2 + (\text{common within-lab variance})$$

Another expression for the left hand-side of this equation is

[33] As an alternative to the Jack-Knife method, simulation may be performed to estimate $SE(\log(SE(\hat{\theta}_i))$. In the early stage of this work the simulation-alternative was used estimate $SE(\log(SE(\hat{\theta}_i))$ in another context as mentioned in the preceding footnote. The relative ease in computing the standard error estimates using the Jack-Knife method holds irrespective of the intended use of them.

$$\mathrm{var}(\hat{\theta}_{\mathrm{i}}) = SD(\hat{\theta}_{\mathrm{i}})^2 = \frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}} \quad \text{.....2.21}$$

(The notation is somewhat imprecise and arbitrary here. In this equation, $\mathrm{var}(\hat{\theta}_{\mathrm{i}})$ denotes variance of $\hat{\theta}_i$ as a random variable while $SE(\hat{\theta}_i)$ on the left hand-side is referring to the analytical standard error of observed $\hat{\theta}_i$. As such, in this equation $\mathrm{var}(\hat{\theta}_{\mathrm{i}})$ and $SE(\hat{\theta}_i)^2$ mean two separate entities.)

The "common within-lab variance" can be derived by subtraction as

$$(\text{common within-lab variance}) = \mathrm{var}(\hat{\theta}_{\mathrm{i}}) - \hat{\tau}^2 = \frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}} - \hat{\tau}^2 \quad \text{.....2.22}$$

or

$$SD_{\text{within-lab}}(\hat{\theta}_{\mathrm{i}}) = \sqrt{\left(\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}} \right) - \hat{\tau}^2} \quad \text{.....2.23}$$[34]

This subtraction method does not generate an estimate of standard error of $SD_{\text{within-lab}}(\hat{\theta}_{\mathrm{i}})$.[35] On the other hand, the direct pooling described in the previous

[34] Dr. Feder states that before pooling within-lab SE's we need to determine whether they are constant across labs It they are approximately constant, a simple arithmetic mean of lab-specific would serve as a good pooled estimate. It is implied that if not constant, pooling is not meaningful since there is no common level to begin with. This argument makes sense, but it seems even if $SE(\hat{\theta}_i)$ are heterogeneous, we can think of a typical level around which the heterogeneous SE's are distributed. The subtraction method computes an estimate of such a measure of central tendency if in an oversimplified manner. DL method is widely-used to obtain estimate of intrinsic between-group variance $\hat{\tau}^2$ and variance of a pooled mean, which combines contribution of $\hat{\tau}^2$ and within-group variance. My understanding is that DL method does not strictly require constant within-group variance across groups. (Its use in meta-analysis that combines trials of vastly different sample sizes seems to be based on lack of such requirement.) As such, irrespective of whether the within-lab variability is constant across labs or not, it seems justifiable to use the subtraction method to compute "average" contribution of within-lab variability to overall between-lab variability. When there is a more justifiable methods exists, at least the idea behind the subtraction method could be used to explain to non-statistician audience how the more justifiable method works as a refined method to achieve the same goal of estimating "typical" level of within-lab variability. In fact, I think the utility of the DL method, which arguably is not the best method as Dr. Feder shows, to EDSP or anybody contemplating the issue of analysis of data from receptor binding assay and performance criteria for the assay is the underlying concepts of the two components (intrinsic between-group and within-group variabilities) in the overall between-group variability.

paragraphs does generate an estimate of the standard error of the pooled mean of $\log(SE(\hat{\theta}_i))$, which may be used to compute an estimate of standard error of $SD_{\text{within-lab}}(\hat{\theta}_i)$. In our applications, we do not need any measure of spread for $SD_{\text{within-lab}}(\hat{\theta}_i)$: $SD_{\text{within-lab}}(\hat{\theta}_i)$ is one of the summary measures of variation that should be reported in an assay validation, but the standard error of $SD_{\text{within-lab}}(\hat{\theta}_i)$ is not one of them. A merit of the subtraction method is that it uses the statistics available as standard output of the curve-fitting procedures, eliminating the need to compute $SE(\log(SE(\hat{\theta}_i))$. The subtraction method was used in "3 Examples"

2.2.4 Jack-Knife variance estimation[36]

As mentioned earlier, in deriving precision criteria and computing an estimate of within-unit variation common to multiple units, a Jack-Knife method may be used in estimating the SE of log(SE(lab-specific logIC_{50})) or log(SE(lab-specific logRBA)).[37]

---

[35] For the simple purpose of computing an estimate of $SD_{\text{within-lab}}(\hat{\theta}_i)$, we do not need its standard error. When we try to analyze multiple $SD_{\text{within-lab}}(\hat{\theta}_i)$ estimates further, the standard error of $SD_{\text{within-lab}}(\hat{\theta}_i)$ may be needed. For instance, let's say that multiple $SD_{\text{within-lab}}(\hat{\theta}_i)$ for different chemicals are available. We may need to summarize these multiple estimates across chemicals or see if $SD_{\text{within-lab}}(\hat{\theta}_i)$ is higher or lower for any particular chemical(s). In another instance, we may have two groups of labs, well-trained and minimally-trained, and would like to compare if the additional training provided for the well-trained labs decreased $SD_{\text{within-lab}}(\hat{\theta}_i)$. In these instances, the standard error of $SD_{\text{within-lab}}(\hat{\theta}_i)$ would be useful. Nonetheless, for our general purposes we do not need the standard error of $SD_{\text{within-lab}}(\hat{\theta}_i)$.
Commenting on the sentence above, Dr. Feder stated "I disagree. You need the SE if you are going to use it as a performance criterion." EDSP has not intended to set up any performance criterion targeted at $SD_{\text{within-lab}}(\hat{\theta}_i)$, which is a pooled estimate for multiple labs in the context of this section. Performance criteria that EDSP has been developing all are intended to be used to determine acceptability of a single lab, not that of multiple labs as a set. I doubt this would change in the future. As such, for the EDSP's "general purposes," On the other hand, standard error of log($SE(\hat{\theta}_{\text{R for lab i}})$) are indeed necessary for setting up precision performance criteria. I have just used $SD_{\text{within-lab}}(\hat{\theta}_i)$ as referred in Equation 2.20. $SD_{\text{within-lab}}(\hat{\theta}_i)$ is a non-pooled version of $SD_{\text{within-lab}}(\hat{\theta}_i)$-like quantity computed for individual lab, not for multiple labs.

[36] There probably is an argument that the boot-strap method of variance estimation is superior to the Jack-Knife method. The Jack-Knife is recommended since its implementation is transparent and there is no need for special statistical software.

[37] Dr. Feder has two comments on this. The first is that delta method is a better procedure to use. The second is "Similarly [to the previously explained solution to "pooling" within-lab SE's] if SE's are the same across labs then use simple average. If they are not the same then why pool?" I do not understand his objection to "pooling" the SE's. This objection seems to contradict his statement ("You need the SE if you are going to use it as a performance criterion.") recorded in the footnote 35 above. Indeed, the "pooling" is necessary for proper description of between-lab variability in log(SE(lab-specific logIC_{50})) or log(SE(lab-specific logRBA)), which in turn is necessary for setting performance criteria for these variability measures. His handwritten suggestion recorded near the end of this section, as I understood it, recommends that we use delta method to compute SE(log(SE(lab-specific logIC_{50}))) or SE(log(SE(lab-specific logRBA))). I take

In general, Jack-Knife variance estimation is performed as follows. Let $\hat{\theta}$ denote the value of the statistic of interest obtained using the entire data set, and let $\hat{\theta}_{(j)}$ denote the value of the same statistic obtained with the *j*th observation omitted. We calculate the "pseudovalue" $\hat{\theta}_j^*$ for the *j*th observation as,

$$\hat{\theta}_j^* = N\hat{\theta} - (N-1)\hat{\theta}_{(j)} \quad \text{.....2.24}$$

where *N* is the total number of observations. The standard error of $\hat{\theta}$ is estimated by taking the standard error of the mean of *N* pseudovalues, i.e., $\hat{\theta}_1^*$, $\hat{\theta}_2^*$, $\hat{\theta}_3^*$, …, $\hat{\theta}_N^*$. $\hat{\theta}$ itself is used as the estimator of the parameter that the statistic of interest estimates. An alternative estimator of the parameter (i.e., the mean of the pseudovalues) is available but its use was not adopted.

When computing the Jack-Knife standard error for a lab-specific summary and the number of runs (where "run" is the observation unit) is greater than 3, the correction for the number of runs performed would be necessary. Such correction has been discussed in "2.2.2 Correction for the number of runs performed".

Let's focus on the SE log(SE(lab-specific logRBA)) for the rest of this section. The same procedure applies to the SE of log(SE(lab-specific logIC_{50})). In estimating the SE of log(SE(*lab*-specific logRBA)), each log(SE(*run*-specific logRBA)) may be treated as an "observation" for which the above-mentioned pseudovalue is computed and summarized across observations in the Jack-Knife method. That is, three log(SE(*lab*-specific logRBA) pseudovalues are computed based on results either from runs "1 and 2", "2 and 3", or "3 and 1" using the DL random effects model. The SE of the mean of these three pseudovalues is the Jack-Knife SE of log(SE(*lab*-specific logRBA).

2.3 Method for evaluating within- and between-replicate-set variation of binding measurements

2.3.1 Definition of unit used in partitioning within- and between-variation

As discussed earlier, we derived upper limits for within-replicate-set and between-replicate-set variabilities of individual binding measurements y at a given x. In other words in doing so we defined the triplicates at any given concentration within a run of a binding assay as a unit of observation.[38]

---

it to mean that those SE's can be used to describe the distribution of log(SE(lab-specific logIC_{50})) or log(SE(lab-specific logRBA)), which in turn can be used to set up precision performance criteria, and he is actually not objecting to the general strategy for setting precision performance criteria.

[38] Why we chose the within-replicate-set and between-replicate-set variabilities as measures to partition the total variability of y given x into "within-run"-like and "between-run"-like variabilities may not be self-evident. There is an alternative to the "replicate set" as a unit used for partitioning such two variabilities.

There is a standard method for partitioning within-unit and between-unit variation called "random effects one-way ANOVA". A typical situation where this method is used involves taking measurements for many subjects (i.e., persons) in duplicate or triplicate. In this setting, measurement error alone is manifested as within-subject variation, which contributes to overall observed between-subject variation. Intrinsic between-subject variation (often expressed as $SD_{between\text{-}subject}$ or a variant such as $SD_{group\ effect}$) is estimated by obtaining overall between-subject variation and subtracting from it the contribution of within-subject variation. If there are 3 subjects with triplicate measurements, the data would look like the following with 9 (= 3 times 3) observations.

Table 2.1 Simple between-subject data structure

	Subject								
	1			2			3		
	Replicate			Replicate			Replicate		
	1	2	3	1	2	3	1	2	3
Observations	y_{11}	y_{12}	y_{13}	y_{21}	y_{22}	y_{23}	y_{31}	y_{32}	y_{33}

In random effects one-way ANOVA, the expected level of within-subject variation is assumed to be the same for all subjects and a single estimate of within-subject variation is obtained.

An extension of this kind of standard random effects one-way ANOVA can be used to define and derive within-replicate-set and between-replicate-set variabilities for data from receptor binding experiments, which take the general form shown in Table 2.2. In this table actual values of x, e.g., -11, -10, etc., were replaced with a generic expression in a form of x_i , i = 1, 2, …, 7. In this table, a simpler notation for y without subscripts is used. In a more precise notation comparable to the one in Table 2.1, subscripts for run, x, replicate would be added, e.g., y_{111} for the cell at the upper left hand corner.

We chose not to use this alternative, but it is described in "4.1 Alternative definition of a unit in partitioning variation in % binding".

Table 2.2 Receptor binding data seen as an extension of between-subject structure

	Run								
	1			2			3		
	Replicate			Replicate			Replicate		
X	1	2	3	1	2	3	1	2	3
x_1	y	y	y	y	y	y	y	y	y
x_2	y	y	y	y	y	y	y	y	y
x_3	y	y	y	y	y	y	y	y	y
x_4	y	y	y	y	y	y	y	y	y
x_5	y	y	y	y	y	y	y	y	y
x_6	y	y	y	y	y	y	y	y	y
x_7	y	y	y	y	y	y	y	y	y

~~Focus on a single row of Table 2.2 and compare that to Table 2.1 for the triplicate data for 3 subjects. They have similar structure. If we are interested in within-replicate-set and between-replicate-set variability at $x = x_1$ only, for instance, we can use the standard random effects one-way ANOVA to estimate such variabilities.~~[39]

~~We would like to estimate the variabilities for the entire data set, encompassing multiple levels of x (or rows), not just a single level of x (or a row). In doing so, we would like to ignore the part of the variation in y attributable to varying levels of x. Thus we need to remove any x-related variation before we can fit a random effects one-way ANOVA model. The variation in y due to x is removed by calculating the mean of y at each x across all runs (see Table 2.3) and subtracting it from each y at that x as illustrated in Table 2.4.~~

[39] Dr. Feder pointed out this setup is incorrect. A correct way to partition variance using two way mixed effects ANOVA is given on page 56 of Feder (2007b). Because of this, the description in this and next section are largely incorrect. To indicate this, affected portion of the texts are shown with strikethrough effect.

Table 2.3 Removal of effects of varying concentration – calculation of x-specific overall mean

	Run									
	1			2			3			
	Replicate			Replicate			Replicate			
x	1	2	3	1	2	3	1	2	3	Mean(y)
x_1	y	y	y	y	y	y	y	y	y	$\bar{y}_1$
x_2	y	y	y	y	y	y	y	y	y	$\bar{y}_2$
x_3	y	y	y	y	y	y	y	y	y	$\bar{y}_3$
x_4	y	y	y	y	y	y	y	y	y	$\bar{y}_4$
x_5	y	y	y	y	y	y	y	y	y	$\bar{y}_5$
x_6	y	y	y	y	y	y	y	y	y	$\bar{y}_6$
x_7	y	y	y	y	y	y	y	y	y	$\bar{y}_7$

That is, we calculate $y^* = y - \bar{y}_{x_i}$ and analyze y^* using random-effects one-way ANOVA treating each combination of x and run as a different unit. In the example above, there are 21 units (7 x levels times 3 runs) across which between-unit variability is estimated. Each unit has 3 observations. A unit as defined here is a replicate set consisting of triplicates at any given log(concentration), or x.

Table 2.4 Modification of receptor binding data for computation of within-unit and between-unit variance: unit specification by run-x combination

	Run								
	1			2			3		
	Replicate			Replicate			Replicate		
x	1	2	3	1	2	3	1	2	3
x_1	$y - \bar{y}_1$	$y - \bar{y}_1$	$y - \bar{y}_1$	$y - \bar{y}_1$	$y - \bar{y}_1$	$y - \bar{y}_1$	$y - \bar{y}_1$	$y - \bar{y}_1$	$y - \bar{y}_1$
x_2	$y - \bar{y}_2$	$y - \bar{y}_2$	$y - \bar{y}_2$	$y - \bar{y}_2$	$y - \bar{y}_2$	$y - \bar{y}_2$	$y - \bar{y}_2$	$y - \bar{y}_2$	$y - \bar{y}_2$
x_3	$y - \bar{y}_3$	$y - \bar{y}_3$	$y - \bar{y}_3$	$y - \bar{y}_3$	$y - \bar{y}_3$	$y - \bar{y}_3$	$y - \bar{y}_3$	$y - \bar{y}_3$	$y - \bar{y}_3$
x_4	$y - \bar{y}_4$	$y - \bar{y}_4$	$y - \bar{y}_4$	$y - \bar{y}_4$	$y - \bar{y}_4$	$y - \bar{y}_4$	$y - \bar{y}_4$	$y - \bar{y}_4$	$y - \bar{y}_4$
x_5	$y - \bar{y}_5$	$y - \bar{y}_5$	$y - \bar{y}_5$	$y - \bar{y}_5$	$y - \bar{y}_5$	$y - \bar{y}_5$	$y - \bar{y}_5$	$y - \bar{y}_5$	$y - \bar{y}_5$
x_6	$y - \bar{y}_6$	$y - \bar{y}_6$	$y - \bar{y}_6$	$y - \bar{y}_6$	$y - \bar{y}_6$	$y - \bar{y}_6$	$y - \bar{y}_6$	$y - \bar{y}_6$	$y - \bar{y}_6$
x_7	$y - \bar{y}_7$	$y - \bar{y}_7$	$y - \bar{y}_7$	$y - \bar{y}_7$	$y - \bar{y}_7$	$y - \bar{y}_7$	$y - \bar{y}_7$	$y - \bar{y}_7$	$y - \bar{y}_7$

This way of defining a unit, which is equivalent to a "subject" in the previous simple situation, may not be self-evident. Nonetheless, this definition of unit appears to be used

commonly. We call the within-unit variability measure computed using this definition of unit "within-replicate-set variation" and express it as $SD_{within\ replicate\ set}$.[40]

Complementary to this within-replicate-set variation is between-replicate-set variation. The within-replicate-set variation does not include any variation due to the effect of x (because by definition of "replicate" x is constant within a replicate set). Since y^* does not include influence of x, between-replicate-set variation computed on y^* also does not include influence of x although the descriptor "between-replicate-set" may indicate otherwise. To be precise, the between-replicate-set variation in y^* probably should be termed "between-replicate-set variation in y given x". For simplicity, though, we call it within-replicate-set variability and express it as $SD_{between\ replicate\ set}$.

2.3.2 Random effects one-way ANOVA

Random effects one-way ANOVA is used to compute $SD_{within\ replicate\ set}$ and $SD_{between\ replicate\ set}$. For the default data setup consisting of 3 runs each having 7 log(concentration) levels (i.e., x levels), there are a total of 21 replicate sets.

Since we regard a replicate set (each combination of run and concentration) as a unit, let i denote each replicate set, $i = 1, 2, \ldots, k$ (where k represents the total number of replicate sets, or 21 in our default setup). Let's denote each replication by j, $j = 1, 2, 3$ for our default setup of triplicate measurements.

Let's denote each observation of $y^* = y - \bar{y}_{x_i}$ as defined in the previous section as y_{ij}. Now we are ready to apply the standard random effects one-way ANOVA to y_{ij}. Using y_{ij}, two mean squares are computed as follows. The nomenclature involving y_{ij} and the procedures described hereafter through the end of this section are standard for random effects one-way ANOVA.

$$MS_\alpha = \sum_i w_i.(\bar{y}_i - \bar{y})^2/(k-1) \quad \text{.....2.25}$$[41]

$$MS_e = \sum_i \sum_j w_{ij}(y_{ij} - \bar{y}_i)^2/(N-k) \quad \text{.....2.26}$$

where

[40] In some documentation that has been presented to EDSP as a part of the method development work to date, this variability measure has been described under different names. The most frequently used was $SD_{within\text{-}run}$, which was used, e.g., in an Excel spreadsheet titled "lab$run_chem@YYMMDDSDw.xls". Explanations for abbreviations used in this spreadsheet appear in "2.3.3 Excel spreadsheet for variability of y within a replicate set".

[41] Dr. Feder pointed out that definition for w_{ij} is missing. For our purposes, $w_{ij} = 1$.

$$w_i = \sum_j w_{ij} \quad \text{.....2.27}$$

$$w = \sum_i w_i \quad \text{.....2.28}$$

$$\bar{y}_i = \sum_j w_{ij} y_{ij} / w_i \quad \text{.....2.29}$$

$$\bar{y} = \sum_i w_i \bar{y}_i / w \quad \text{.....2.30}$$

N is the number of observations and k is the number of units.

The standard deviation within the replicate set, σ_ε or $SD_{\text{within replicate set}}$, is estimated as

$$\sigma_\varepsilon = \sqrt{MS_\varepsilon} \quad \text{.....2.31}$$

The standard deviation of unit effects, σ_α or $SD_{\text{between replicate set}}$, is estimated as

$$\sigma_\alpha = \sqrt{(MS_\alpha - MS_\varepsilon)/g} \quad \text{.....2.32}$$

where

$$g = \frac{w - \sum_i w_i^2 / w}{k-1} \quad \text{.....2.33}$$

For the case of equal unit size and no missing data, g = "number of observations in each unit" = N/k.

2.3.3 Excel spreadsheet for variability of y within a replicate set

An Excel spreadsheet titled "lab$run_chem@YYMMDDSDw.xls" computes within-replicate-set variation in y for data from a single run. Explanations for abbreviations used in the "Calculation" datasheet in this Excel file are given below.

x: log(concentration). [42]

y: Observed % binding values.

x_with_y: This column shows the value of x only when y on the same row was non-missing.

[42] The font used for the abbreviations is matched to the one that was used in the Excel spreadsheet titled "lab$run_chem@YYMMDDSDw.xls" for easy comparison.

my_x: Mean of y given x. There usually are three y observations for each level of x. When there is a missing y (e.g., one y observation missing/removed), it should calculate the mean based on duplicate measurements rather than the usual triplicate measurements. This quantity corresponds to $\bar{y}_i$ in ~~Table 2.3~~ and ~~Table 2.4~~.

yij: Individual y value minus mean of y given x, or y - my_x. This quantity corresponds to y_{ij} as defined in "2.3.2 ~~Random effects one-way ANOVA~~".

myio: Mean(y) at *i*th level of x or

$$\bar{y}_i = \sum_j w_{ij} y_{ij} / w_i \quad \text{.....2.29}$$

as defined in "2.3.2 ~~Random effects one-way ANOVA~~". In this spreadsheet, it is zero since y_{ij} is defined as the original y value - (mean(y) at *i*th level of x).

yijMNyiosq: Square of (individual y value minus the mean(y) at *i*th level of x). This quantity corresponds to $(y_{ij} - \bar{y}_i)^2$ as it appears earlier in [43]

$$MS_\varepsilon = \sum_i \sum_j w_{ij} (y_{ij} - \bar{y}_i)^2 / (N - k) \quad \text{.....2.26}$$

sum_yijMNyiosq: Sum of squares of "the individual y value minus the mean(y) at *i*th level of x". This quantity corresponds to $\sum_i \sum_j w_{ij} (y_{ij} - \bar{y}_i)^2$ in the equation for MS_ε above.

N: Number of observations as used in the equation above.

k: Number of x levels as used in the equation for MS_ε above.

Mse: Mean within-unit sum of squares (MS_ε). See the MS_ε equation above.

2.4 Method for describing the relationship between variability measures[44]

As stated in the section "1.4.3.2 Performance criteria for within-replicate-set variability of y", the upper limit for within-replicate-set variation of y is derived by finding the level

[43] Although the originally proposed ANOVA model had incorrect degrees of freedom for the between-replicate-set effects as pointed out by Dr. Feder, the formula for the within-replicate-set variance was correct.

[44] Additional explanation regarding why the general approach described in this section was originally favored by EDSP is given in Aoki (2007b). Over time this initial preference by EDSP has been lost, and I recommended that EDSP forgo the originally developed method and use a method comparable to the one used for setting performance criteria for Hill equation parameters in the revision of performance criteria prompted by Dr. Feder's feedback. As such, materials in this section now has limited relevance to EDSP's performance criteria development activity. Because of this and because of limited time available for me to revise this report, I am keeping response to Dr. Feder's comments in this section at minimum.

of within-replicate-set variation of y corresponding to the SE(lab-specific $\log IC_{50}$ estimate) upper limit. To do this, we need a quantitative description of the bivariate relationship between the within-replicate-set variation of y and SE(lab-specific $\log IC_{50}$ estimate).

Intuitively, in data for many runs collected under a common experimental design for a given analyte, there should be a relationship between the overall between-replicate-set variation and the precision of the estimated $\log IC_{50}$. It may be possible to derive this relationship on a theoretical basis, but such derivation is beyond the scope of this report. The discussion below deals only with empirically determining such a relationship.

The relationship was not expected to be an exact deterministic kind since there would be unexplained variation. (This could be confirmed easily in the actual interlab study data by plotting the data.) Thus, some kind of regression method is called for. For reasons to be explained later, ordinary least square regression is used to describe the quantitative relationship. This relationship was then used to derive the performance criteria for within-replicate-set variability of y.

There is a certain intuitive basis for the existence of a monotonically increasing relationship between the two variability measures in question. For a given design (a pre-determined number of x levels surrounding the target $\log IC_{50}$ with appropriate intervals so that y would reach both top and bottom plateaus within the range of x levels), the greater the overall between-replicate-set variation is, the greater the SE(lab-specific $\log IC_{50}$ estimate) should be.

By analogy to simple linear regression, this intuition would hold if the following condition holds since the mean sum of squares based on the left-hand side is an estimate of overall between-replicate-set variance of y given x and the square root of the residual mean square based on the right-hand side is proportionate to the standard error of the estimated parameters.

$$E(Y \mid X = x) = E\left(Y \mid Y = \hat{B} + \frac{\hat{T} - \hat{B}}{1 + 10^{[(\hat{\mu} - X) * \hat{H} + \log(\frac{\hat{T} - \hat{B}}{50 - \hat{B}} - 1)]}} + \varepsilon, X = x \right) \quad \text{.....2.34}$$

(Note that the expectation on the left hand-side involves no parameterized model.)

In assessing the bivariate relationship, it is natural to choose the dependent (outcome) and independent (predictor) variables as

Dependent variable ≡ overall between-replicate-set variability in y given x
Independent variable ≡ SE(lab-specific $\log IC_{50}$ estimate)

since the level of the former corresponding to a certain level of the latter (i.e., the upper limit value) is the quantity of interest.

The estimate of overall between-replicate-set variability in y given x may be obtained only after data from multiple (at least 2) runs are collected. EPA was also interested in limiting the level of within-replicate-set variation in y, which can be assessed with data from a single run, and is contemplating use of criteria imposed on it. There are various conceptual approaches for how to set up such criteria, and a method based on the bivariate relationship between within-replicate-set variability in y and SE(lab-specific $\log IC_{50}$) was eventually chosen.

The within-replicate-set variability in y is one of two components of the overall between-replicate-set variability in y given x (the other component being the intrinsic between-replicate-set variability). Upon assessment of the RUC interlab data, a positive association between the within-replicate-set variability in y and SE(lab-specific $\log IC_{50}$) was found. As mentioned earlier, there is a reason to believe the within-replicate-set variability is directly associated with the SE(lab-specific $\log IC_{50}$). In addition to that direct association, there presumably was an indirect association between the within-replicate-set variation and SE(lab-specific specific $\log IC_{50}$) due to positive correlation between within-replicate-set variability and between-replicate-set variability, both of which reflect the quality in execution of laboratory work. For instance, the precision of pipetting technique can affect both within- and between-replicate-set variability.

A positive correlation also was found between the intrinsic between-replicate-set variability in y given x and SE(lab-specific $\log IC_{50}$). As expected, a positive correlation between the overall between-replicate-set variability in y given x and SE(lab-specific $\log IC_{50}$) also was found. This correlation was slightly weaker than that between the intrinsic between-replicate-set variability and SE(lab-specific $\log IC_{50}$). This was unexpected since intuitively the overall between-replicate-set variability in y given x was thought to most closely reflect the underlying overall noise level in y, which is proportional to SE(lab-specific specific $\log IC_{50}$). The discrepancy probably was due to random variation as we have only 8 data points (four laboratories each for two chemicals) for the regression involving overall or intrinsic between-replicate-set variability in y as a dependent variable.[45]

In general, the bivariate relationship between independent variable U and predictor V is described in the following general form

$$U = \beta_0 + \beta_1 V + \varepsilon \quad \text{.....2.35}$$

where ε is an error term. Estimates of β_0 and β_1, $\hat{\beta}_0$ and $\hat{\beta}_1$, are obtained by fitting an appropriate model to the data. The rest of this section describes certain features that need to be exercised in choosing an appropriate model for description of bivariate relationships.

Scale: It is desirable to analyze the logarithms of both the dependent variable and the independent variable because both of them are non-negative by definition (and almost

[45] When the designations for independent and dependent variables were reversed and log(SE(lab-specific $\log IC_{50}$)) was regressed on $\log(SD_{\text{total-between-replicate-set}})$ and $\log(SD_{\text{intrinsic-between-replicate-set}})$, $\log(SD_{\text{total-between-replicate-set}})$ appeared to be a better predictor of log(SE(lab-specific $\log IC_{50}$)) than $\log(SD_{\text{intrinsic-between-replicate-set}})$.

certainly positive in any realistic situations we encounter). The non-negative nature of both of these implies their relationship is multiplicative (consider the fact that any predicted decrease beyond zero is meaningless), and computation performed on an absolute scale may produce nonsensical results such as a negative standard deviation. Variance or standard deviation could be used as a measure of quantities on either side of the regression equation. Once a logarithm is taken these two measures are equally appropriate as long as the same measure is used on both sides of the regression equation since:

$$\text{variance} = (\text{standard deviation})2 \rightarrow \log(\text{variance}) = 2 \times \log(\text{standard deviation}) \quad .....2.36$$

and multiplication by two on both sides of the regression equation cancels out each other.

Within-lab correlation: Within-replicate-set variability estimates from the same laboratory are correlated to each other when multiple laboratories are under consideration.[46] However, independence among observations typically is assumed in certain regression techniques, including the ordinary least square (OLS) regression method. Correlation is allowed in more robust methods developed as an extension of the standard method. Such robust methods would need to be used in order to assess the strength of evidence for a bivariate association in a receptor-binding dataset. As long as only point estimates of bivariate regression are concerned, however, the OLS and its robust counterpart generate the same results.[47] The OLS regression also provides a readily implemented option for computing a predictive band around the fitted line, which we need when deriving the upper limit for the dependent variable. Although the use of the OLS means we are ignoring the within-lab correlation, from the inspection of the predictive band computed it was felt the consequence of ignoring this correlation was not serious.

2.5 Deriving accuracy criteria

2.5.1 $\log IC_{50}$, Top, Bottom, Hill slope

A common method may be used for deriving accuracy criteria for the four parameters $\log IC_{50}$, top, bottom, and Hill slope. Following the notation used in the section "2.2 Method for between-run and between-lab summary of parameters", a parameter is referred to generically as θ in this section.

---

[46] Dr. Feder questions "Why do you have multiple within-replicate-set variability for the same lab?" Isn't within-replicate-set variability pooled over multiple runs within lab?" Because of the aforementioned interest of EDSP to set limit for within-replicate-set variability measured after each run, the analysis were performed using data set of log(SDwthin-replicate-set) computed for each run and lab-specific SE($\log IC_{50}$). (That is, all log(SDwthin-replicate-set) from the same lab had a common lab-specific SE($\log IC_{50}$).
In addition to the check for within-replicate-set variability for each run, EDSP also once planned to check within-replicate-set variability for three runs together. The description of the method for calculating within-replicate-set variability in this document was intended for such "3 runs together" analysis, but I failed to clearly document these background nor the within-replicate-set variability estimation method based on data from a single run. EDSP's original plan regarding how these checks were to be performed is illustrated in a slide titled "How criteria will be used" on page 15 of Aoki (2007b).

[47] Dr. Feder comments "No. You will get different point estimates, sometimes very different estimate." This statement does not agree with actual computational result I had. I applied either OLS or its robust version to the data set involving estimates of log(SDwithin-replicate-set) and log(SE($\log IC_{50}$)) and obtained exactly the same point estimates with different standard errors.

To derive accuracy criteria, we first need to select a set of laboratories that are deemed acceptable.[48] After analyzing the data from an initial interlab study, these labs were chosen by informal judgment, which is more of a practical, rather than a scientific, nature. From a scientific standpoint, it is best to accept laboratories that meet a very strict standard. It may be infeasible to do so because of the cost for training a laboratory and other practical obstacles. From an administrative standpoint, laboratories should be as good as possible within the practical constraints such as the cost and feasibility of finding laboratories meeting the performance criteria. Past experience of the EDSP in terms of locating laboratories and ensuring a certain quality of data from them was deemed to be relevant in making an administrative judgment as to what "as good as possible within practical feasibility" means. In the revision of performance criteria after a round of subsequent interlab study, there would be various approaches to the selection of acceptable labs. An approach would be to define acceptable labs as the labs participating in the subsequent interlab study that meet the provisional performance criteria based on the original interlab study. This seems most straightforward. It might be justifiable, though, to "fortify" the set of new acceptable labs with the "deemed-to-be-acceptable" labs in the original study if the number of the new acceptable labs is small.[49] Also, any "novice" labs meeting the provisional performance criteria, whether it is in the original or subsequent study, may be chosen as acceptable.[50] Either way, the initial selection of acceptable laboratories drives the whole process of criteria derivation.

Let's say we have selected *k* acceptable laboratories. We are concerned about the distribution of $\hat{\theta}$, which is an estimate of θ to be reported from a laboratory drawn from a universe of laboratories like the ones that are deemed acceptable. As described in the section "2.2.1 DerSimonian-Laird random effects model", standard deviation of $\hat{\theta}$ is computed as

$$SD(\hat{\theta}) = \sqrt{\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}} \quad \text{.....2.18}$$

This measure of spread includes two sources of variation, within-lab and between-lab. To appreciate this, consider two extreme cases.

[48] Dr. Feder states "This is extremely important. Width of prediction intervals depend on this." I agree. There are different philosophies as to how this is done conceptually, which I tried to depict in Section "2.3.2. Prediction interval vs. tolerance interval" of Aoki (2007b). I believe EDSP and Dr. Feder are operating under different philosophy.

[49] If we take the "fortification" approach for the interlaboratory study that follows the existing interlab study for the RUC ER binding assay, it would be advisable to include "novice" labs only and exclude lab E, which was the leading lab.

[50] Not all "deemed-to-be-acceptable" labs would meet the provisional performance criteria as will be shown in Chapter 3 "Examples".

If $\hat{\tau}$ is zero (that is, intrinsic between-lab variation is zero) and $SE(\hat{\theta}_1) = SE(\hat{\theta}_2) = ,... = SE(\hat{\theta}_k)$ (that is, the standard errors of the parameter estimates within each single laboratory are equal), then the formula simplifies to

$$SD(\hat{\theta}) = SE(\hat{\theta}_i) \quad \text{.....2.37}$$

that is, the standard deviation of the parameter estimate across laboratories is also the standard error of the parameter estimate within a single laboratory, implying that the within-lab variation is the sole contributor to the distribution of parameter values across labs.

If $SE(\hat{\theta}_1) = SE(\hat{\theta}_2) = ,... = SE(\hat{\theta}_k) = 0$, (that is, the standard errors of the parameter estimates in each laboratory are all zero) then the formula becomes

$$SD(\hat{\theta}) = \hat{\tau} \quad \text{.....2.38}$$

That is, the variability in the parameter is due only to between-lab variation.

Using the pooled estimate of within-lab variance described in "2.2.3 Pooled estimation of within-lab variability", the formula for $SD(\hat{\theta})$ may be rewritten as

$$Var_{\text{overall}}(\hat{\theta}) = \hat{\tau}^2 + \left(\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}} - \hat{\tau}^2 \right) \quad \text{.....2.39}$$

where $\left(\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}} - \hat{\tau}^2 \right)$ is the average within-lab variation.[51] This formula has the following generic form introduced in the first chapter:

Overall (total) between-unit variability = intrinsic between-unit variability + overall within-unit variability

Using this estimator of the spread and the estimator of the central location, we can construct a range for $\hat{\theta}$ within which, at a specified probability, an estimate of θ to be reported from a laboratory drawn from the universe of laboratories like the ones that are deemed acceptable will fall. Assuming Normality, such a range (prediction interval) at probability coverage of (1-α)*100% is computed as

[51] The equation might look somewhat odd as $\hat{\tau}^2$ appears twice on the right side, and they appear to cancel each other out. This is a regularly used technique to get at a quantity one would like.

$$\hat{\theta}_R \pm t_{1-\alpha/2,\,\nu} * \sqrt{1+1/k} * SD(\hat{\theta}) \quad \text{.....2.40}$$

where ν is Hartung and Makambi (2001) degrees of freedom and, as presented earlier,

$$\hat{\theta}_R = \frac{\sum_{i=1}^{k} w_i^* \hat{\theta}_i}{\sum_{i=1}^{k} w_i^*} \quad \text{.....2.14}$$

$$SD(\hat{\theta}) = \sqrt{\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}} \quad \text{.....2.18}$$

$$w_i^* = \frac{1}{\hat{v}_i + \hat{\tau}^2} \quad \text{.....2.13}$$

For instance, an 80% prediction interval is computed as

$$\hat{\theta}_R \pm t_{0\,90,\,\nu} * \sqrt{1+1/k} * SD(\hat{\theta}) \quad \text{.....2.41}$$

$t_{1-\alpha/2,\,\nu}$ instead of $t_{1-\alpha,\,\nu}$ is used since $t_{1-\alpha/2,\,\nu}$ corresponds to (1-α/2)*100 percentile of Normal distribution and we are interested in the range that encompasses from α/2*100 percentile to (1-α/2)*100 percentile.

2.5.1.1 Analytical vs. simulation-based SE(lab-specific logIC$_{50}$) and SE(lab-specific logRBA)

As described in "2.8.2.1 Comparing methods for summarizing data from multiple runs", simulation studies were performed. It was found that the analytical standard errors of SE(lab-specific logIC$_{50}$) and SE(lab-specific logRBA)—that is, the within-lab variability of the logIC$_{50}$ estimate and that of the logRBA estimate —computed using the DerSimonian-Laird analytical method appeared to underestimate the realistic level of within-lab variability, which may be approximated by simulation. The degree of underestimation was quantified and a correction factor was applied to the observed analytical standard error to derive more realistic standard error values. However, whether we use the analytical standard error as is or with the correction, there is little difference in the derived accuracy criteria for logIC$_{50}$ and logRBA. Detail of this will be presented in "3.5.6 Impact of underestimation of SE(lab-specific logIC$_{50}$) and SE(lab-specific logRBA)". Since it will be cumbersome to perform simulation every time we need to use analytical SE(lab-specific logIC$_{50}$) and the consequence of ignoring the underestimation seems small, it seems justifiable to the analytical SE(lab-specific logIC$_{50}$) and SE(lab-specific logRBA) without correction on a routine basis.

2.5.2 logRBA

Relative binding affinity is defined as

$$\text{RBA} \equiv (\text{IC50 of standard}) / (\text{IC50 of test chemical}) \quad \text{.....2.42}$$

Accordingly, logRBA is defined as

$$\text{logRBA} \equiv (\text{logIC}_{50} \text{ of standard}) - (\text{logIC}_{50} \text{ of test chemical}) \quad \text{.....2.43}$$

SE(logRBA) is determined assuming that the (logIC_{50} of the standard) and the (logIC_{50} of test chemical) are independently distributed. However, the standard is concurrently run with a test chemical and the (logIC_{50} of standard) is computed from data for the concurrently-run standard, so the assumption of independence between (logIC_{50} of standard) and (logIC_{50} of test chemical) probably does not hold in truth. In general, the variance of logRBA is given as

$$\begin{aligned}&\text{Var(logRBA)}\\&= \text{Var}(\text{logIC50}_{\text{standard}}) + \text{Var}(\text{logIC50}_{\text{test chemical}}) - 2 \cdot \text{Cov}(\text{logIC50}_{\text{standard}}, \text{logIC50}_{\text{test chemical}})\end{aligned} \quad \text{.....2.44}$$

For lack of an easy way to quantify the covariance, we ignore the covariance term.[52] By the independence assumption, variances are converted to standard errors by taking the square root of both sides.

$$\text{SE(logRBA)} = \sqrt{\text{SE}(\text{logIC}_{50} \text{ of standard})^2 + \text{SE}(\text{logIC}_{50} \text{ of test chemical})^2} \quad \text{.....2.45}$$

Once pairs of logRBA and SE(logRBA) estimates are obtained for each run of a chemical, results are summarized across runs and then across labs by fitting the DL random effects model. A prediction interval of the same form as that for other parameters, i.e., $\hat{\theta}_{\text{R}} \pm \text{t}_{1-\alpha/2,\nu} * \sqrt{1+1/k} * SD(\hat{\theta})$, is then computed.

2.6 Setting precision criteria

2.6.1 Standard error of logIC_{50}

$\text{SE(logIC}_{50}\text{)}$ is non-negative and $\log(\text{SE(logIC}_{50}\text{)}$ would be more symmetrically distributed than $\text{SE(logIC}_{50}\text{)}$ itself. Working with $\log(\text{SE(logIC}_{50}\text{)}$ rather than $\text{SE(logIC}_{50}\text{)}$ is desirable for these reasons. By choosing to do so, we are avoiding the occurrence of a nonsensical, negative lower limit of a prediction interval for a standard error, which might arise when working with $\text{SE(logIC}_{50}\text{)}$.[53]

[52] Dr. Feder states "With enough runs you can estimate Var(logRBA) directly." I can see that would true for between-run Var(logRBA). In this section, though, I was referring to within-run Var(logRBA), which can be used as input for between-run summary of logRBA using DL method. If there is an easy way to compute within-run Var(logRBA), it could have been used in Feder and Ma (2005), but apparently such a method was not used in Feder and Ma (2005).

[53] Dr. Feder notes "Also, the distribution is more nearly symmetrically thereby giving better approximation to asymptotic theory in small samples."

Derivation of precision criteria for SE($logIC_{50}$) starts with selection of acceptable laboratories in terms of SE($logIC_{50}$). This is analogous to the initial step in the derivation of accuracy criteria. The set of laboratories with acceptable between-lab SE($logIC_{50}$) does not need to be the same as the set having acceptable $logIC_{50}$ estimates or other parameter estimates. The same caveat about the informal, practical nature of this selection applies here.

In order to apply the DL random effects model to log(SE($logIC_{50}$)), we need SE(log(SE($logIC_{50}$))), which was estimated using simulation in the initial effort to set up performance criteria. The simulation was performed by Battelle and described fully in its report (Feder and Ma, 2005). Alternatively, it may be computed using the Jack-Knife method described in the section "2.2.4 Jack-Knife variance estimation". In future performance setting exercises, the Jack-Knife method would be a preferable approach because of its simplicity.[54]

Once estimates of log(SE($logIC_{50}$)) and SE(log(SE($logIC_{50}$))) are computed, we proceed by treating them as $\hat{\theta}_i$ and $SE(\hat{\theta}_i)$ and use the method described in the section "2.5.1 $logIC_{50}$, Top, Bottom, Hill slope" to derive $\hat{\theta}_R$ and $SD(\hat{\theta})$.

Since low levels of SE($logIC_{50}$) are not problematic (it actually is desirable to have low SE($logIC_{50}$)), we do not need to set a lower limit for SE($logIC_{50}$). For SE($logIC_{50}$) we set up an upper limit only, in the following form.

$$\hat{\theta}_R + t_{1-\alpha/2,\nu} * \sqrt{1+1/k} * SD(\hat{\theta}) \quad \text{.....2.46}$$

For instance, the upper limit for an 80% prediction interval is set as follows.

$$\hat{\theta}_R + t_{0\,80,\nu} * \sqrt{1+1/k} * SD(\hat{\theta}) \quad \text{.....2.47}$$

An estimator of $SD(\hat{\theta})$, which uses information available as standard output of the DL random effects model applied to $\hat{\theta}_i$ (that is, lab-specific parameter estimates) is

$$SD(\hat{\theta}) = \sqrt{\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}} \quad \text{.....2.48}$$

This could be rewritten as

[54] See footnote 15 for preference given to delta method over Jack-Knife.

$$SD_{\text{overall}}(\hat{\theta}) = \sqrt{\hat{\tau}^2 + \left(\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}} - \hat{\tau}^2 \right)} = \sqrt{SD_{between}(\hat{\theta})^2 + SD_{within}(\hat{\theta})^2} \quad \text{.....2.49}$$

This formula involves an estimator of within-lab variation in the following form,

$$\sqrt{Var_{within}(\hat{\theta})} = \sqrt{\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}} - \hat{\tau}^2} \quad \text{.....2.50}$$

Instead of this estimator of $SD_{within}(\hat{\theta})$, the other estimator of $SD_{within}(\hat{\theta})$ described in "2.2.3 Pooled estimation of within-lab variability" may be used in computing $SD_{\text{overall}}(\hat{\theta})$. This estimator was based on a "DL random effects model"-based overall mean of log(SE($\hat{\theta}$)). As shorthand, let the following be a formula for such an overall mean and its standard error.

$$\left(\hat{\theta}_{\text{R}}\right)_{\log(SE(\hat{\theta}_i))} = \left(\frac{\sum_{i=1}^{k} w_i^* \hat{\theta}_i}{\sum_{i=1}^{k} w_i^*} \right)_{\log(SE(\hat{\theta}_i))} \quad \text{.....2.51}$$

$$\left(SE(\hat{\theta}_{\text{R}})\right)_{\log(SE(\hat{\theta}_i))} = \left(\sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}} \right)_{\log(SE(\hat{\theta}_i))} \quad \text{.....2.52}$$

Using this shorthand, an alternative estimator of $SD_{within}(\hat{\theta})$ has the following expression.

$$SD_{\text{within, direct pooling}}(\hat{\theta}) = 10^{\left(\hat{\theta}_{\text{R}}\right)_{\log(SE(\hat{\theta}_i))} + \left(SE(\hat{\theta}_{\text{R}})\right)_{\log(SE(\hat{\theta}_i))}^{2} / \log(e)} \quad \text{.....2.53}$$

(Derivation of this involving the formulae to convert the mean and variance for original and log-transformed lognormal random variable is omitted.)[55]

Plugging this into $SD_{\text{overall}}(\hat{\theta}) = \sqrt{SD_{between}(\hat{\theta})^2 + SD_{within}(\hat{\theta})^2}$, i.e., the formula for the overall between-lab variation, an alternative estimator of $SD_{\text{overall}}(\hat{\theta})$ is

[55] Dr. Feder recommends the use of "$\sqrt{\chi^2_{\nu/\nu}}$ for SE distribution".

$$SD_{\text{overall}}(\hat{\theta}) = \sqrt{\hat{\tau}^2 + (10^{\left(\hat{\theta}_{\text{R}}\right)_{\log(SE(\hat{\theta}_i))} + \left(SE(\hat{\theta}_{\text{R}})\right)^2_{\log(SE(\hat{\theta}_i))}/\log(e)})^2} \quad \text{.....2.54}$$

Whether this estimator performs favorably compared to $SD(\hat{\theta}) = \sqrt{\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}}$ or not has not been investigated.

2.6.2 Standard error of logRBA

The procedure used to set precision criteria for logRBA is virtually the same as that for $\log IC_{50}$.

2.6.3 Within- and between-replicate-set variability of binding measurements

Derivation of criteria for within- and between-replicate-set variability of individual y measurements or y given x includes two components: limits established for the overall within-lab variation of $\log IC_{50}$ and; relationship between the overall within-lab variation of $\log IC_{50}$ and the specific variability measure of individual y measurements.

The performance criteria that have been developed to date include criteria for within-replicate-set variability of y given x. Let $UL(\text{variability measure})$ denote an upper limit derived for a particular measure of variability. Also, suppose there is the following relationship between the two variability measures of interest.

$$\log(SD_{within-replicate-set}(\text{y})) = \beta_0 + \beta_1 \log(\text{SD}(\hat{\theta})) + \varepsilon \quad \text{.....2.55}$$

as a particular example of

$$\text{U} = \beta_0 + \beta_1 \text{V} + \varepsilon \quad \text{.....2.35}$$

Estimates of β_0 and β_1 (i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$) may be obtained by fitting an appropriate model to the data. The following equation holds for $UL(SD_{within-replicate-set}(\text{y}))$, which is the upper limit for $SD_{within-replicate-set}(\text{y})$.

$$\log(UL(SD_{within-replicate-set}(\text{y}))) = \hat{\beta}_0 + \hat{\beta}_1 \log(UL(\text{SD}(\hat{\theta})) + \text{STDF} * Z_{0\,95} \quad \text{.....2.56}$$

where STDF is standard error of the forecast, which combines the error in prediction and residual error. Exponentiating both sides, the upper limit for $SD_{within-replicate-set}(y)$ is computed as

$$UL(SD_{within-replicate-set}(y))=10^{\hat{\beta}_0+\hat{\beta}_1 \log(UL(\hat{SD}(\hat{\theta}))+STDF *Z_{0.95}}$$

Upper limits for the overall between-replicate-set variability and intrinsic between-replicate-set variability in y given x, may be computed in a similar manner by describing the bivariate relationship between the measure of interest and $SD(\hat{\theta})$.

2.7 Relative utility of accuracy and precision criteria

Accuracy criteria for top, bottom, and Hill slope ensure the general quality of data. The accuracy criteria for $logIC_{50}$ of the standard are of limited use since even among competent labs $logIC_{50}$ varied considerably.

An accuracy criterion for logRBA of a positive control chemical is more useful than an accuracy criterion for $logIC_{50}$ of the reference chemical or positive control.

The precision criteria for $logIC_{50}$, logRBA, and within- and between-replicate-set variability of individual measurements given x are, like the accuracy criteria for the three "shape" parameters, useful for ensuring internal consistency of data.

2.8 Justification for some assumptions and solutions

Certain supporting analyses for the assumptions are summarized in this section.

2.8.1 Constant noise assumption

The results of the analyses that support the operating assumption of constant noise level are presented in this section. The analysis investigated the noise levels for one historical set of RUC ER binding assay data.

For estradiol data, k was estimated to be 0.13, indicating the error structure was more like a Normal distribution than a Poisson distribution. For norethynodrel (the positive control chemical), the estimated k was 0.26, giving more—albeit not overwhelming—support for a Poisson distribution than for the constant error structure.

The residuals for the above analysis were calculated using unweighted nonlinear regression. Using a weight of 1/Y would give rise to a different set of parameter estimates, predicted values, and residuals. It is possible that residuals from a 1/Y-weight regression show Poisson-like structure. For this reason, the entire analysis was repeated using 1/Y as a weight. This alternative analysis yielded k-estimates of 0.11 and 0.31 for estradiol and norethynodrel, respectively, meaning that again for estradiol, the constant error level was supported; and for norethynodrel the Poisson error structure was slightly more supported.

Transformation of the outcome variable can "stabilize" error structures (i.e., make the error structure of the transformed variable Normal). In other words, after transformation, the error becomes independent of the outcome variable. For Poisson error structure, square root transformation is such a variance-stabilizing transformation. Since our outcome variable sometimes takes a negative value, we can use $f(Y) = sign(Y)\sqrt{|Y|}$ as a variance-stabilizing transformation and change the model equation accordingly as follows:[56]

$$sign(Y)\sqrt{|Y|} = sign\left(B + \frac{T-B}{1+10^{(\mu-X)*H+\log\left(\frac{T-B}{50-B}-1\right)}}\right)\sqrt{\left|B + \frac{T-B}{1+10^{(\mu-X)*H+\log\left(\frac{T-B}{50-B}-1\right)}}\right|} + \varepsilon \quad \text{.....2.57}$$

The idea of a variance-stabilizing transformation provides another way to investigate which error structure is most supported by the data. For instance, if Poisson distribution is truly the underlying error structure, the mean squared residuals of the transformed variable regressed on X using the modified model equation would be constant across the transformed variable. This premise was examined using the RUC interlaboratory data. The analysis yielded k estimates of -0.17 and -0.15 for estradiol and norethynodrel, respectively, each of which has 95% confidence intervals that exclude 0. These results indicate that the residuals decrease as the average y increases, instead of staying constant as the assumption of Poisson error on the original scale would imply. Therefore, the Poisson error assumption is not supported by the transformed historical data on estradiol or norethynodrel.

Table 2.5 summarizes estimates of k for estradiol, norethynodrel and daidzein obtained using three different fitting procedures.

Table 2.5 Estimated *k* for estradiol, norethynodrel, and daidzein based on three alternative transformation-weighting schemes

Assumed error structure (and correct k value)	Transformation	Weight	Estimated value of k		
			Analyte		
			Estradiol	Norethynodrel	Daidzein
Constant (k = 0)	None	Equal	0.13	0.26	0.32
Poisson (k = 0.5)	None	1/|Y|	0.11	0.31	0.18
Poisson (k = 0)	Sign(Y)*sqrt(|Y|)	Equal	-0.17	-0.15	-0.88

Similar analyses were performed on data collected on another 7 chemicals (Bisphenol B, 4-Cumylphenol, Estrone, Coumestrol, Tamoxifen citrate, 4-tert-Octylphenol, and Bisphenol A). Table 2.6 summarizes results for the data from three labs (C, D, and E), two labs (C and E) or lab D alone. Overall, the data more strongly, if not overwhelmingly, support a homoskedastic error structure than a Poisson error structure when the data were

[56] Dr. Feder objects this procedure, stating "I disagree. Negative values should not get larger weights than small positive values." My attempt to defend this procedure is included in Aoki (2007b).

analyzed without transformation. The Poisson error structure is slightly more favored by the untransformed data from lab D, but when the same data were analyzed with the square root transformation there is much less evidence for a Poisson error structure (k = -0.49). (Analyses by different subsets of laboratories were not performed for estradiol, norethynodrel, and daidzein because of small sample sizes for each of these chemicals.)

Table 2.6 Estimated *k* for various chemicals based on three alternative transformation-weighing schemes

Assumed error structure (and correct k value)	Transformation	Weight	Estimated value of k		
			Laboratory		
			All (C, D, E)	C and E	D
Constant (k = 0)	None	Equal	0.22	0.17	0.27
Poisson (k = 0.5)	None	1/|Y|	0.22	0.17	0.24
Poisson (k = 0)	Sign(Y)*sqrt(|Y|)	Equal	-0.22	-0.17	-0.49

The fact that estimated $k \approx 0.2 > 0$ indicates the error level tends to increase as Y increases, but the rate of increase is not as fast as the level expected for a Poisson error structure. As noted earlier, there are practical obstacles for incorporating a Poisson error assumption in the nonlinear regression curve-fitting procedure. Taken together, the best course of action was determined to be to use the standard weighting option for nonlinear regression, i.e., equal-weighting without transforming the Y variable.

2.8.2 DerSimonian-Laird random effects model

2.8.2.1 Comparing methods for summarizing data from multiple runs

The method discussed in "2.2.1 DerSimonian-Laird random effects model" for summarizing receptor binding data from multiple runs consists of two steps. The first step is to fit the Hill equation to data from each run by nonlinear regression. The second step is to summarize sets of the parameter estimates obtained in the first step by fitting a DerSimonian-Laird random effects model parameter by parameter. An alternative method fits the Hill equation to data sets from multiple runs. It simultaneously provides run-specific summaries and between-run summaries for the four parameters of the Hill equation, complete with estimates of between-run variation for each of the four parameters. A modification of this procedure also may be used to estimate logRBA by fitting a properly-parameterized model to a combined data set from concurrently-performed runs for a standard and test chemical.

In a simulation study conducted by Battelle (Feder and Ma, 2005), this alternative method was compared to the method based on the DL random effects model. The goal of this study was to investigate the appropriateness of using the DL random effects model for summarizing parameter estimates from multiple runs. The alternative curve-fitting procedure available as PROC NLMIX in SAS that simultaneously produces summary statistics across runs was found to be impractical because of the issue of non-convergence and very lengthy computation time.

Comparison of the analytical standard error for a parameter estimate from fitting a DL random effects model with the observed standard deviation of estimates of the parameter

from simulation revealed that the standard error was underestimated by the DL procedure, but not by much. The maximum likelihood (ML) alternative to the DL procedure, which is a moment-based method, was found to perform better than the DL in that the standard error was less biased. The ML procedure involves iterative estimation and is much less accessible compared to the DL method. That is, the DL method involves closed-form solutions only and can be performed manually or by using a spreadsheet program while the ML procedure is available only in SAS and other costly commercial software.

Based on the results of the simulation study and concerns for accessibility, it was judged that the practical advantage of the DL procedure prevails over its statistical shortcomings.[57]

2.8.2.2 Underestimation of between-unit variability in the conventional method

In "1.3.2.2 Limitations in the conventional approach to estimating variability between laboratories" it was claimed that the conventional method for estimating between-unit variability, which relies solely on the observed overall between-lab variability, ignoring the estimated within-unit variability, tends to underestimate the true level of overall between-lab variability. This tendency is more pronounced when the true intrinsic between-unit variability is small compared to the true within-unit variability.

These claims are based on simulations. The simulations performed, along with their results, are summarized below.

In each round of simulation, three realized values of θ_i , $i = 1, 2, 3$, are drawn from a Normal distribution with mean zero and variance k^2, $N(0, k^2)$. For each of these, $\hat{\theta}_i$ were generated as a mean of l realized random variable values. Each one of these l values is generated as,

$$\theta_{ij} = \theta_i + \delta_j \ , \ \delta_j \sim N(0,1) \ , \ j \ = \ 1, \ 2, \ ..., \ l \qquad \text{.....2.58}$$

i.e., with Normally-distributed random error such that the expectation of $SE(\hat{\theta}_i) \equiv 1$.

The choice of l determines the variability of $SE(\hat{\theta}_i)$. The greater l is, the smaller the across-run variability in simulated $SE(\hat{\theta}_i)$ values will be. It was determined that an l of 4 or 5 corresponds to the level of the variability in $SE(\hat{\theta}_i)$ observed in the previous interlab study for the RUC estrogen receptor binding assay. How this conclusion has been drawn is summarized in the next few paragraphs.

The observed levels of between-run variability (standard deviation) of $\log(SE(\hat{\theta}_i))$ are shown in Table 2.7 for the combination of laboratories (A, C, D, E) and analytes

[57] As noted earlier, the preference to the more accessible procedure is largely of EDSP, not of me.

(estradiol used as standard and norethynodrel used as positive control) for the $\log IC_{50}$ parameter. (Data from the first run for norethynodrel from lab A was dropped because of very poor quality.)

Table 2.7 Observed SD($\log(SE(\hat{\theta}_i))$ **)**

Laboratory	Analyte	
	Estradiol	Norethynodrel
A	0.108	0.189
C	0.146	0.242
D	0.146	0.211
E	0.205	0.081

$SD(\log(SE(\hat{\theta}_i)))$ of about 0.2 seems to be a representative level.

On the other hand, the simulated values of $\log(SE(\hat{\theta}_i))$ generated as described above had the following standard deviations when *l* was varied from 2 to 20.

Table 2.8 Standard deviation of simulated $\log(SE(\hat{\theta}_i)$
corresponding to the number of within unit replications

Value of *l*	2	3	4	5	10	20
$SD(\log(SE(\hat{\theta}_i))$	0.486	0.275	0.213	0.173	0.108	0.071

From this table we can see that the level of *l* corresponding to $SD(\log(SE(\hat{\theta}_i)) \approx 0.2$ is 4~5.

For demonstration purposes, an *l* of 5 was chosen. Using the simulated data, we can compute two separate estimates of $SE(\hat{\theta}_R)$, one using the conventional method of simply taking the mean of three run-specific summaries and calculating its standard error and the other using the DL random effects model. The former ignores $SE(\hat{\theta}_i)$ associated with each $\hat{\theta}_i$—that is, it ignores within-run variability. The latter incorporates $SE(\hat{\theta}_i)$ in the computation. Let's call these $SE(\hat{\theta}_R)_{conv}$ and $SE(\hat{\theta}_R)_{DL}$.

In an extreme case where the ratio of intrinsic between-run SD to within-run SD (*k*) is 0, i.e., the intrinsic between-run variation is zero (or there is no between-run heterogeneity), the simulated distributions of $SE(\hat{\theta}_R)_{conv}$ and $SE(\hat{\theta}_R)_{DL}$ are shown in Figure 2.3. The vertical line indicates the level of $SE(\hat{\theta}_R)$ expected from between-run and within-run variabilities specified in the simulation.

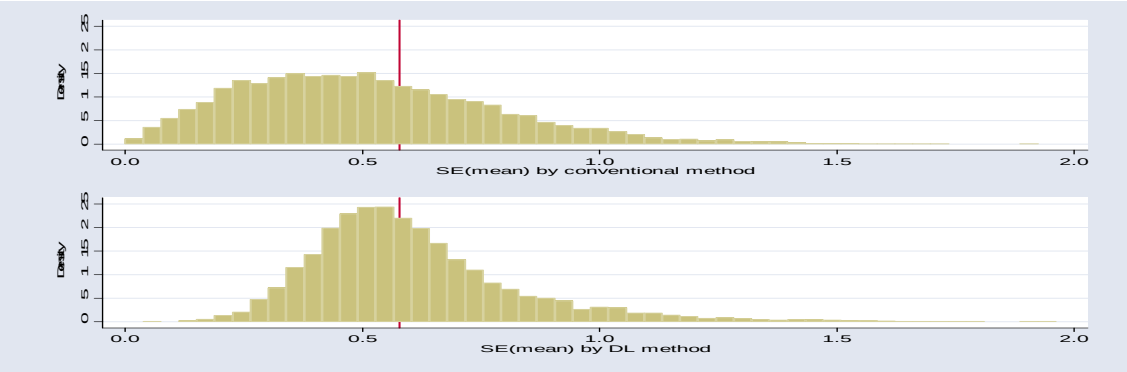

Figure 2.3 Comparison of $SE(\hat{\theta}_R)_{conv}$ and $SE(\hat{\theta}_R)_{DL}$, intrinsic between-run variability = 0

Note that the distribution for $SE(\hat{\theta}_R)_{conv}$ has greater probability mass near zero away from the expected center or its center is shifted towards zero from its expected center than that for $SE(\hat{\theta}_R)_{DL}$. This indicates that the chances of underestimating the true level of overall between-run variability are greater for the conventional method than for the DL method. This happens because values of $\hat{\theta}_i$ may happen to be close to each other, making $SE(\hat{\theta}_R)_{conv}$ small. When this happens, a user of the conventional approach is stuck with an unusually low estimate of $SE(\hat{\theta}_R)$. In the same situation, the DL method computes the overall between-run variation by adding the observed between-run variation, which would be at or near zero, and within-run variation, which most of time is NOT near zero.

The degree of underestimation depends on the relative contribution of the intrinsic between-run variation and within-run variation to the overall between-run variation. As the contribution of the intrinsic between-run variation increases, the tendency of the conventional method to underestimate the overall between-run variation diminishes. In order to get a good understanding of the realistic level of this underestimation, *k* (the hypothesized ratio of intrinsic between-run SD to within-run SD) needs to be set to a realistic value.

In the data from the previous interlab study, the observed ratios of intrinsic between-run SD to within-run SD were computed as follows.[58]

[58] The ratios shown in Table 2.9 differ from those presented in Table 3.3 and Table 3.4. They are different because the two set of ratios were computed in a different manner. In Table 2.9, the arithmetic mean of the ratio values computed for each lab-analyte combination is shown. In contrast, in Table 3.3 and Table 3.4, a ratio was computed for each lab-analyte combination by dividing the intrinsic between-run SD by the pooled estimate of within-run SD. Since the ratio tended to be right-skewed, a ratio values in Table 2.9 is greater than the corresponding ration in Table 3.3 and Table 3.4.

Table 2.9 Observed ratio of $SD_{\text{intrinsic between-run variability}}$ to $SD_{\text{within-run variability}}$

Laboratory	Analyte	
	Estradiol	Norethynodrel
A	1.7	1.1
C	3.6	4.2
D	3.7	2.4
E	1.7	0.6

There is a single estimate of the between-run SD for each lab-analyte combination, and the ratio is computed taking this common between-run SD and each of the run-specific within-run SD estimates (the number of runs ranges from 3 to 12). In order to show the run-to-run variation of this ratio, the estimated run-specific ratios are summarized using box plots in Figure 2.4.

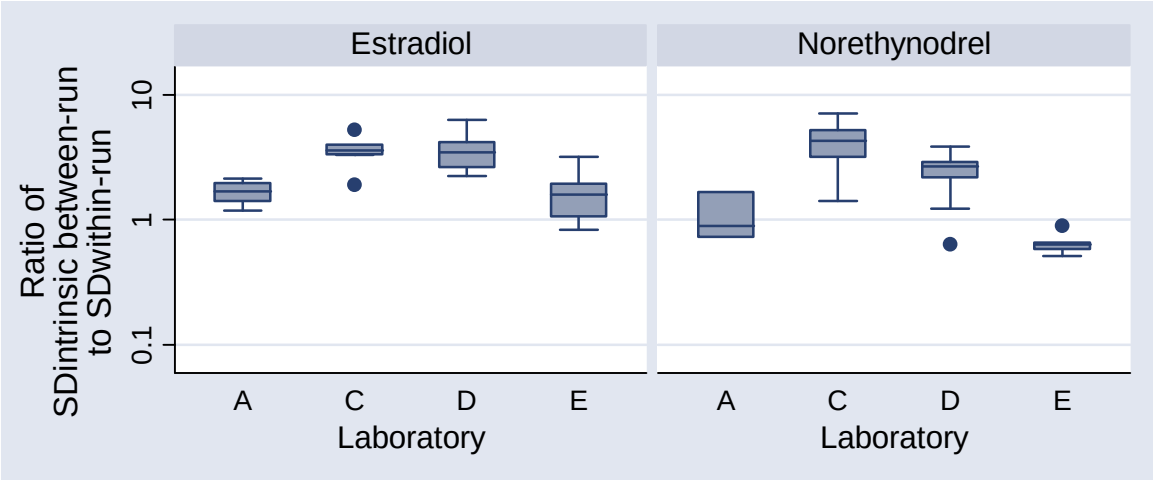

Figure 2.4 Observed ratios of $SD_{\text{intrinsic between-run}}$ to $SD_{\text{within-run}}$ by laboratory

The average "between-to-within" ratio most of the time was greater than 1 for labs C and D. A value of the ratio near or below 1 occurs on occasion for lab A. Ratios close to or below 1 were not uncommon for lab E, and in fact lab E's ratios for norethynodrel were all below 1. Note that lab E was the lab whose performance was the best among the 4 labs. Lab E had the lowest overall between-run variability largely due to its low intrinsic between-run variability.

Figure 2.5 below shows a comparison of simulated distributions of $SE(\hat{\theta}_R)_{conv}$ and $SE(\hat{\theta}_R)_{DL}$ in the case where the intrinsic between-run is not zero and the between-to-within ratio is one, i.e., $k = 1$.

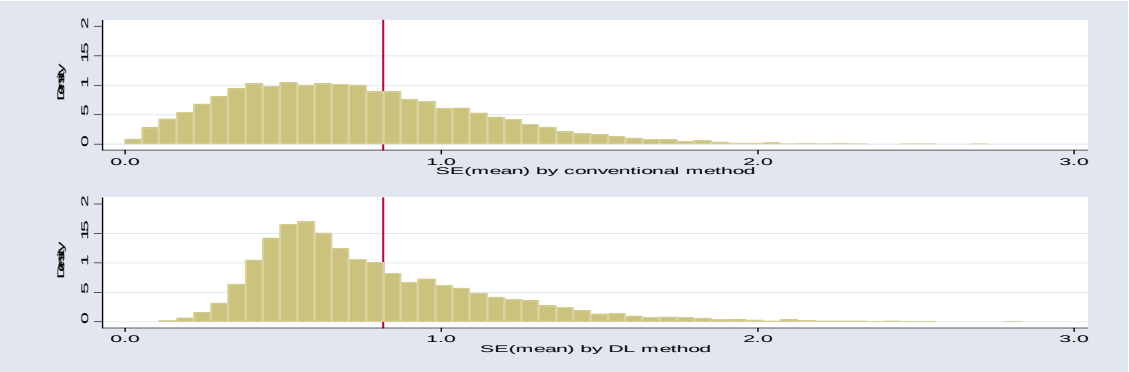

Figure 2.5 Comparison of $SE(\hat{\theta}_R)_{conv}$ and $SE(\hat{\theta}_R)_{DL}$,
$SD_{intrinsic\ between\text{-}run\ variability} = SD_{within\text{-}run\ variability}$

The degree of the underestimation does not seem serious, but it still is noticeable. When *k* is increased to 4, though, the underestimation becomes hardly noticeable as shown in Figure 2.6 below.

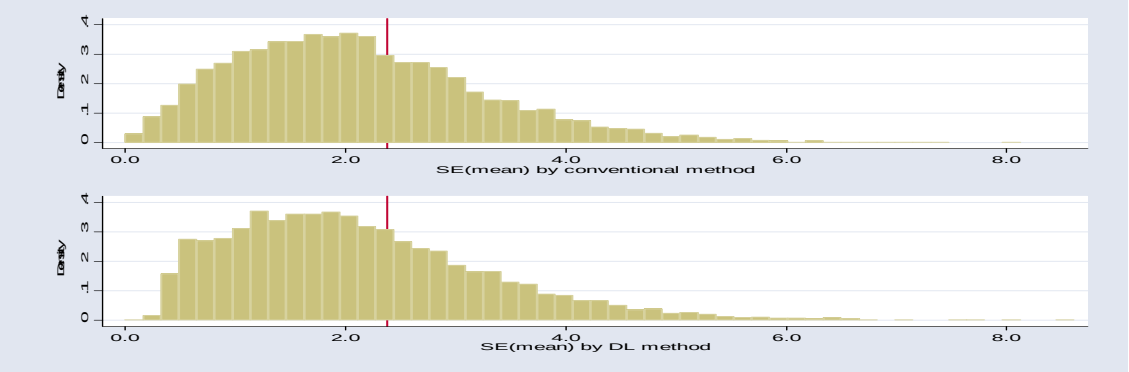

Figure 2.6 Comparison of $SE(\hat{\theta}_R)_{conv}$ and $SE(\hat{\theta}_R)_{DL}$,
$SD_{intrinsic\ between\text{-}run\ variability} = 4 \times SD_{within\text{-}run\ variability}$

Taken together, at certain realistic levels of relative magnitudes of between-run variation and precision of $SE(\hat{\theta}_i)$, $SE(\hat{\theta}_R)_{DL}$ would perform better than $SE(\hat{\theta}_R)_{conv}$.

Let us try a similar analysis for estimation of between-lab variation of $\hat{\theta}$. The observed levels of between-lab variability (standard deviations) of $\log(SE(\hat{\theta}_i))$ were 0.210 for estradiol and 0.374 for norethynodrel, meaning the variability corresponds to an l of 4 for estradiol and an l of 2~3 for norethynodrel for .[59] The observed average ratio of

[59] Unfortunately, the use of $\log IC_{50}$ results as an example here was not very useful because $\log IC_{50}$ estimates are already known to be not comparable across labs. The observed ratios of between-lab variability to within-run variability often were much smaller for other parameters, e.g.:
0 for top plateau parameter (norethynodrel) as shown in Table 3.7;
0 for bottom plateau parameter (both estradiol and norethynodrel) as shown in Table 3.8 and Table 3.9Table 3.10;
0 for slope parameter (estradiol) as shown in Table 3.10.

intrinsic between-lab SD to within-lab SD was 4.4 for estradiol and 8.1 for norethynodrel.[60] These relatively high ratios indicate that the intrinsic between-lab variation is the predominant source of variation in the overall between-lab variability of $\log IC_{50}$ estimates. It is unlikely that $SE(\hat{\theta}_R)_{DL}$ and $SE(\hat{\theta}_R)_{conv}$ generate diverging estimates of the overall between-lab variabilities due to the above-discussed underestimation associated with $SE(\hat{\theta}_R)_{conv}$. Note, though, that such assurance becomes available to us only if we look at the relative contribution of the intrinsic between-lab variability to the overall between-lab variability, about which the DL method can inform us but the conventional method cannot. As long as we rely on the conventional method alone, we would have no way of assessing whether the potential underestimation was of real concern or not. Interestingly, the above results indicate that as an assay improves in terms of reduced intrinsic between-lab variability, the likelihood of underestimating the overall between-lab variability increases.

2.9 Other uses of variability estimates

The use of separate estimates of the intrinsic between-unit and within-unit variabilities for setting performance criteria has been described. There are other uses of them, some of which are described below.

2.9.1 Uses in assay development

Separate estimates of the intrinsic between-unit and within-unit variabilities may be useful at a pre-validation stage when one is trying to identify specific sources of variation as targets of efforts to reduce overall variability. High variability in radioactive count measurement, for example, would tend to increase within-run variation, not intrinsic between-run variation. Inappropriate preparation of a stock standard solution for each run, from which appropriate serial dilutions can be made reliably, would result in an increase in intrinsic between-run variation, not within-run variation. The fact that such sources of variation are likely to be associated more with one or the other of the two variabilities, combined with knowledge of the observed contribution of these two variabilities, helps us identify specific sources of variation to which our efforts to reduce variation should be targeted.

2.9.2 Uses in assay implementation

For an instance of post-validation use of the between and within variability estimates, suppose the overall between-lab variability for an assay has been found to be unacceptably high under a specified design in all or most of novice laboratories, and we would like to know whether a feasible increase in the number of runs per laboratory would reduce the variability to a meaningful degree and how much of an increase would reduce the variability to the desired level. Only with the separate estimates of intrinsic

---

These represent a situation where underestimation of overall between-lab variability by the conventional method is potentially substantial at least in terms of the "between/within" ratio. The variability of between-lab variation as it affects the precision of $SE(\hat{\theta}_i)$ has not been investigated.

[60] These mean ratio values differ from the "Between/within" ratio values shown in Table 3.3 and Table 3.4 for the reason explained in the footnote 58.

between-lab variability and overall within-lab variability (which is a function of the number of runs), would easy calculation of the necessary number of runs be possible.

Outside of validation exercises, we generally do not have a chemical tested by two or more laboratories. If intrinsic between-lab variability is very high, though, testing by multiple laboratories may be justified.[61] This may happen when, for instance, intrinsic between-lab as well as within-lab variability for logRBA is found to be high for a particular class of chemicals of regulatory importance. The agency may want to control precision of logRBA by requiring testing by multiple laboratories and performing more than three runs per laboratory. The optimal combination in terms of the laboratory cost and gain in precision may be determined using the estimates of intrinsic between-lab and within-lab variabilities. An example of exercise of this kind is given for the combined between-replicate-set variability in individual binding measurement given x in the section "3.6.3.3 Total between-replicate-set variability of % binding of radioligand".

[61] Dr. Feder notes "I disagree. A large component of variability across labs indicate the assay or the protocol needs to be improved. Assay is not reproducible. Sources of variation need to be eliminated. Running greater number of variable labs does not increase reproducibility of assay." Please note that I am recommending this as a routine solution. It would be great if an agency has limitless time and resources to keep improving unsatisfactorily variable assay, but in reality it usually has some timeline and limited resources. In the case of RUC ER binding assay, EDSP felt that it had done as much assay improvement it could perform. It is reasonable to suggest a *potential* solution for the situation where the time (and resources) for assay development has been exhausted but the assay still has unacceptably high between-lab variability. Using multiple labs is shown here as such a last-resort solution. Consider also the following specific situation. Even for assay with acceptably low between-lab, during the implementation of large scale screening an agency may encounter a (newly emerged) class of test chemical for which between-lab variability is found to be much higher than for those tested during assay development.

3 Examples[62]

3.1 Structure of this chapter

This chapter shows how the methods described in Chapters in 1 and 2 were applied to the data collected as a part of the activities called "Task 6 " of Work Assignment 2-30, Contract No. 68-W-01-023, and performance criteria were derived. The title of Task 6 was

Establish the variability of results among the five independent laboratories when using Battelle-supplied "standard" cytosol preparation, 17β-estradiol, and a range of 8-10 test chemicals

A part of the same data was used in the analysis performed by Battelle (Feder and Ma, 2005). The overall structure of this chapter is as follows.

"3.2 Overall organization of the data" is a short section that summarizes how many runs are used for each combination of chemical and laboratory. Correction for the number of runs performed is necessary in order to make results comparable across laboratories.

The following section "3.3 Run- and lab-specific estimates of the Hill equation parameters" provides a graphic overview of the data in terms of run-by-run summary estimates for each laboratory of the Hill equation parameters and laboratory-by-laboratory summary estimates of them. These are computed using the DerSimonian-Laird random effects model. Graphic summary of the data of this kind is useful in making an informal judgment regarding which laboratories are "acceptable" in terms of accuracy and precision of a parameter estimate. This section focuses more on the data rather than the methods. Nonetheless, the usefulness of the DL model as a tool for data description is exemplified in this section.

In the next section titled "3.4 Within-run, between-run (= within-lab), and between-lab variations", an example of actual computation of the pooled means and summary variability measures using the DL method is given as well as an example of actual full numerical summary for an entire interlaboratory study data set in terms of pooled means and various between- and within-unit variability estimates.

The estimates of across-lab pooled mean and overall between-lab variation of the Hill equation parameter described in Section 3.4 are the key quantities in deriving accuracy criteria for those parameters. In "3.5 Deriving accuracy criteria", how to derive accuracy criteria are described for each of the Hill equation parameters. Approaches similar to

[62] EDSP has decided to revise the performance criteria derived in this section. The revision is to be performed separately in part based on Aoki 2007c). To illustrate the originally proposed method, most of description were left unchanged from the version submitted earlier (Aoki, 200a). Some of the necessary changes to the original proposed method, which were suggested by Dr. Feder, are incorporated where possible (in the form of correcting formulae and text), but performance criteria numbers and interim results are left unchanged. As such, performance criteria numbers in this section in general are not to be used.

those used for deriving the accuracy criteria with necessary modifications are used for the derivation of the precision criteria, which is described in the last section "3.6 Setting precision criteria".

Since the EDSP's priority for data quality has changed over time, some potential alternatives to cope with the change were considered and described.

3.2 Overall organization of the data

Table 3.1 gives the number of runs performed for each chemical by laboratory. For most laboratory-chemical combinations there are more than 3 runs. We are interested in a standard condition of three runs performed for a test chemical under the current protocol. The consideration for extra runs described in "2.2.2 Correction for the number of runs performed" applies to the analysis of this data, meaning that when we try to generate a summary across labs we modify the input data so that the results will be relevant to future situations where three runs per chemical are the norm.

Table 3.1 Number of runs for which usable data are available

Laboratory	$logIC_{50}$, top plateau, bottom plateau, and Hill slope for Estradiol	$logIC_{50}$, top plateau, bottom plateau, and Hill slope for Norethynodrel	logRBA for Norethynodrel
A	4	3	3
C	6	6	6
D	7	12	7
E	6	6	6

Usable data on norethynodrel from lab A were available for 3 runs. There was an additional run for this chemical-lab combination, but the data for that run was obviously erratic and excluded from the analysis. Lab B was dropped during the qualification process because of its poor performance.

3.3 Run- and lab-specific estimates of the Hill equation parameters

In order to give an overall visual impression of data variability, so-called forest plots (Lewis and Clarke, 2001) are produced for each of the four parameters: $logIC_{50}$, top, bottom, and slope. Forest plots typically are used in epidemiological publications reporting results of meta-analysis, which is a systematic review of multiple independent studies complete with a quantitative summary of study results.

They are useful for the purpose of this section in presenting results across runs (or labs) because they give visual impressions of within-run (lab) as well as between-run (lab) variability. In general, they are useful for summarizing results from independent "units", whether the units are runs, laboratories, clinical trials, or observational studies.

In the forest plots below, the results from each unit are represented by a rectangle and a horizontal line. The horizontal location of the center of a rectangle corresponds to the point estimates of the parameter of interest from individual units. The area of the rectangle is inversely proportional to the variance of the point estimate. The edges of the horizontal line represent the lower and upper 95% confidence limits for the point estimate. The overall mean of the estimates, as derived by fitting the DL random effects model to them, is shown as the horizontal location of the center of the diamond at the bottom of each plot. Again, the left and right edges of the diamond represent the lower and upper 95% confidence limits for the estimate of the overall mean.

Features of forest plots can be understood easily through the use of an example. Let's take a look at the set of forest plots summarizing the results for $\log IC_{50}$, shown in Figure 3.1. When the results from multiple units are homogeneous or between-unit variability is low, the rectangles are aligned in terms of horizontal location. An example of such a situation is the panel for lab E and norethynodrel at the right bottom corner. When the rectangles are similar in size or the lengths of horizontal lines are similar, that means within-unit variability varied little across units (as is the case, again, with lab E and norethynodrel).

When the intrinsic between-unit variability is high (or results are heterogeneous across units), a plot often includes horizontal lines that exclude the center of another horizontal line or, in more extreme cases, do not overlap with each other. An example of such a situation is seen in the two panels for lab C where the horizontal lines at the top in each panel are away from the rest of the horizontal line. The relatively high heterogeneity for results from lab C is reflected in the high $SD_{\text{intrinsic between-run variability}}$ to $SD_{\text{within-run variability}}$ ratios shown in Table 2.9 "Observed ratio of $SD_{\text{intrinsic between-run variability}}$ to $SD_{\text{within-run variability}}$". Note that some lines in the panels for lab D also exclude each other, and that is consistent with the high values of the ratio for lab D in Table 2.9.

A large difference in the size of rectangles indicates that some point estimates were much more precise than others. In the presence of such a large difference, the DL method is more suitable in producing a more precise estimate of the overall mean than the conventional method. Although this gain in precision is not the main reason for us to use the DL method (the main reason is its capacity to separate within- and between-unit variation), this feature can also be thought of as an additional merit since the precise estimation of overall mean for each lab results in a better estimate of between-lab variation, which is ultimately the quantity of most interest.

By comparing plots for the same chemical across laboratories, we can gain some insights as to whether the results are different across labs. The informal visual impression we receive is that there are differences across laboratories. Lab E appears to give higher $\log IC_{50}$ values than others, and lab D lower $\log IC_{50}$s.

A formal investigation on whether results vary across labs is carried out by fitting the DL random effects model to the lab-specific $\log IC_{50}$s and their standard errors. The result of across-lab summarization is shown in another set of forest plots (Figure 3.2). As a

preparation for producing these, adjustment for the number of runs performed for each lab (see “2.2.1 DerSimonian-Laird random effects model”) was made.

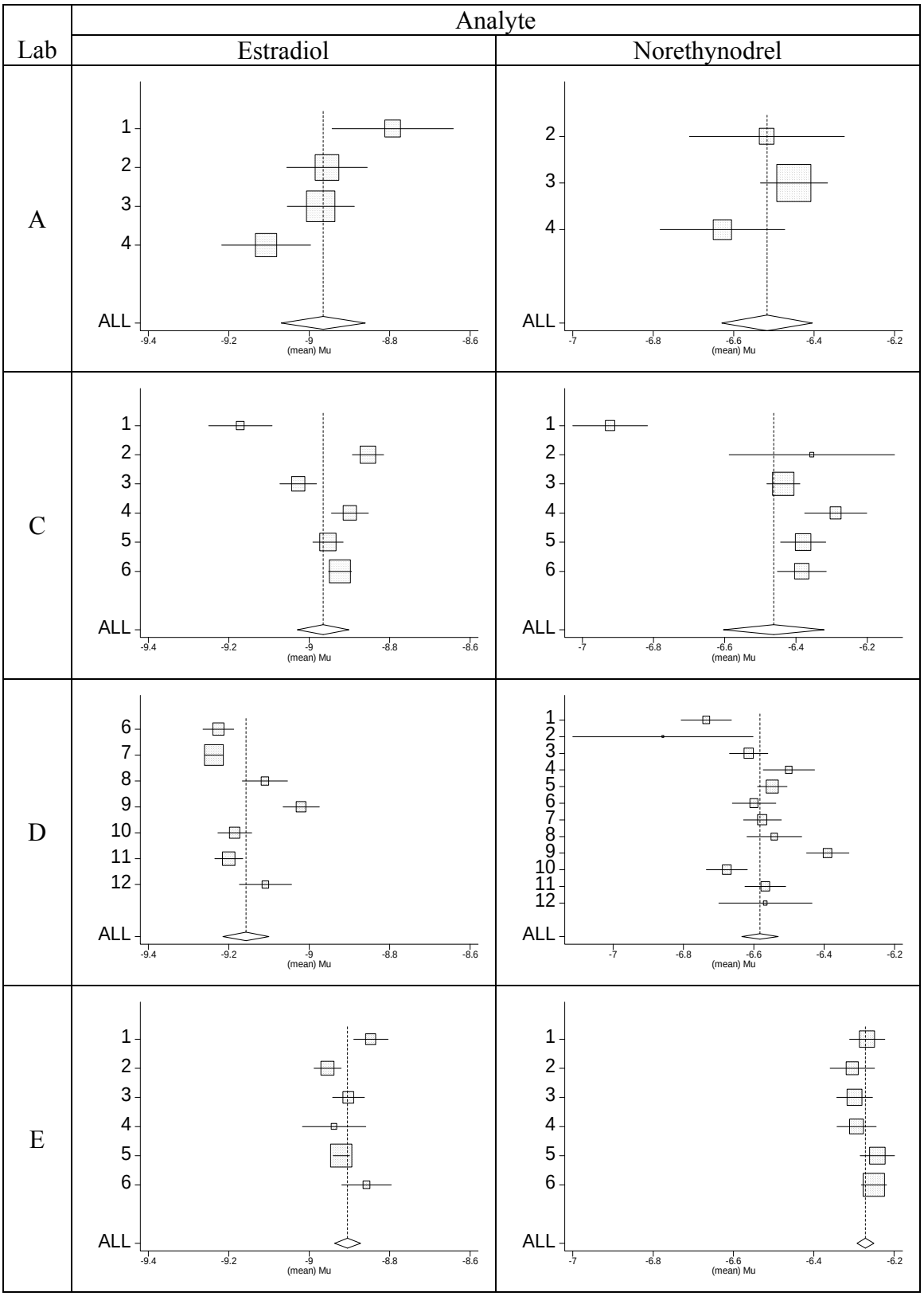

Figure 3.1 logIC_{50} estimates by analyte and laboratory

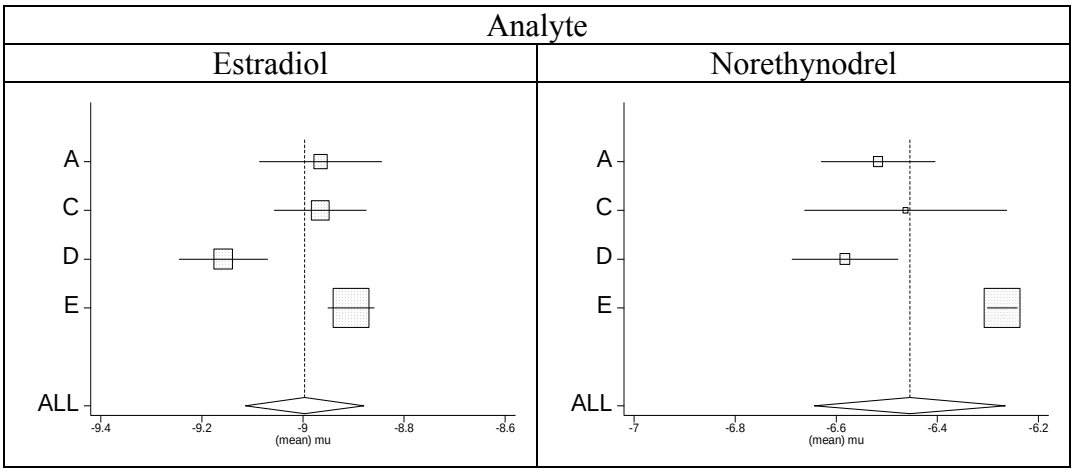

Figure 3.2 Across-lab summary of logIC_{50} estimates

Between-lab heterogeneity is easily noticeable within each panel in Figure 3.2. From a different perspective of comparing the observed pattern across two panels, a parallel shift for logIC_{50} for estradiol and norethynodrel, which implies a type of similarity across chemicals within lab, appears to exist, i.e., deviation of estradiol logIC_{50} for a lab from the overall mean is in the same direction for those of norethynodrel logIC_{50}. Within-lab variation, as represented by the inverse of the area of rectangle and reflected in the length of the horizontal line (i.e., the width of the 95% confidence interval), is the smallest for lab E, indicating relatively good precision by lab E.

This heterogeneity probably is primarily due to variation in the concentration of protein originated in the RUC preparation used. The protein concentration in the RUC preparation can vary considerably across labs. Since the protein concentration has substantial influence on logIC_{50} and standardizing the protein concentration for RUC preparations is infeasible, the EDSP has decided to not impose criteria for accuracy for logIC_{50}.[63]

Some signs of between-lab heterogeneity were observed for the estimates of the other 3 parameters, but they are not as conspicuous as those for logIC_{50}.

[63] In the absence of the knowledge on the variable protein concentration and how strongly it affects logIC_{50}, the same conclusion to not enforce performance criteria for accuracy of logIC_{50} still might have been reached. The observation that the logIC_{50} estimates from lab E appeared different from those from other labs raised concerns as to whether these four laboratories are representative in a general sense. As a background, lab E was the leading lab and had gone through the most intensive scrutiny regarding how experiments were conducted. Information collected by observing actual experimental sessions at lab E was used to improve the written laboratory protocol. The final protocol that underwent such improvement was distributed to laboratories A, C, and D to be used the current interlab study.

It might have been the case that there were certain practices in lab E that were not completely reflected in the final protocol, and the same practices also resulted in its stellar performance in terms of precision as well as its "eccentric" logIC_{50} values. This would mean there was qualitative between-lab heterogeneity in the performance of four laboratories. If this is the case, statistical results based on the combined data from lab E and other labs would not have had good interpretation. The DL random effects model allow heterogeneity across labs, which is quantitative in nature. The potential heterogeneity of qualitative nature hypothesized above is not something the model can properly describe.

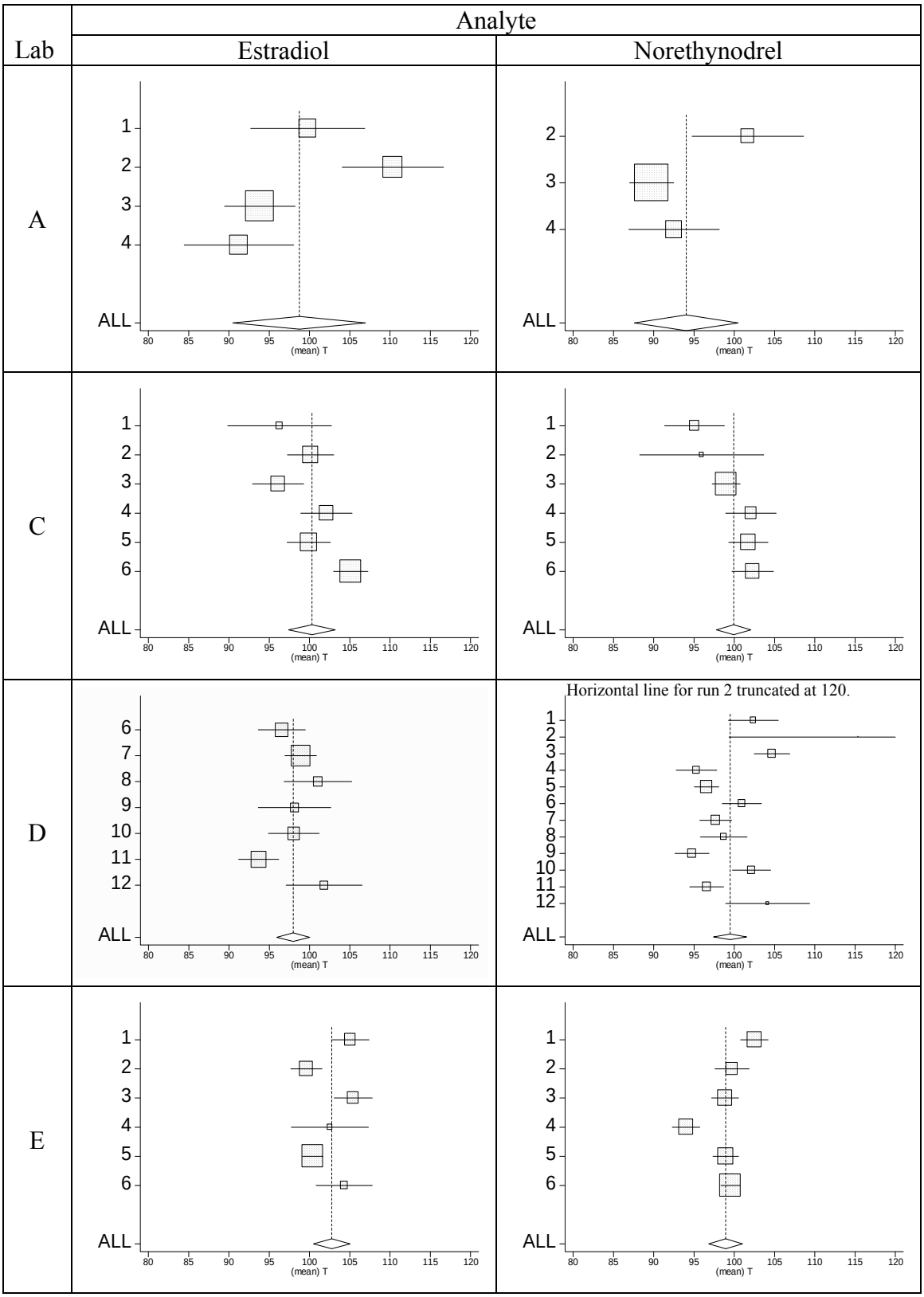

Figure 3.3 Top parameter estimates by analyte and laboratory

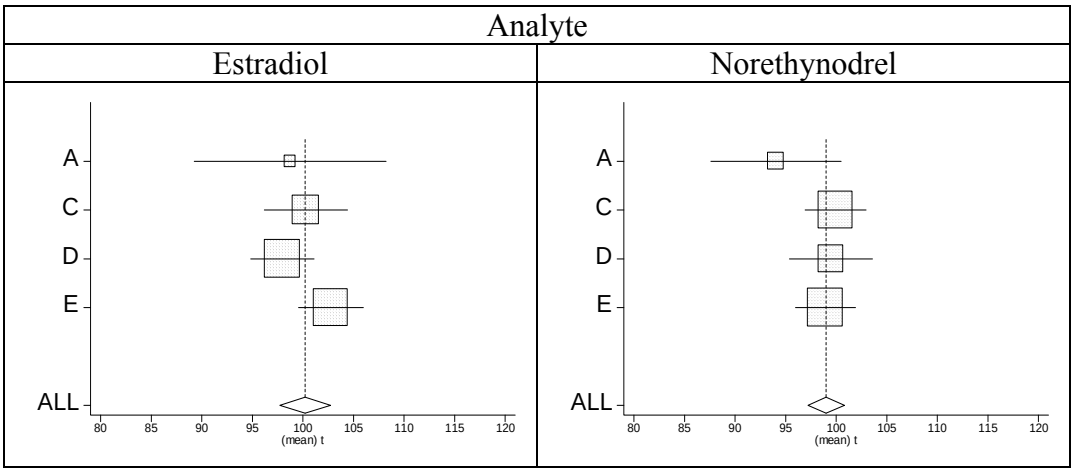

Figure 3.4 Across-lab summary of the estimates of the top plateau level

Within-lab variation for the estimate of the top plateau was higher for lab A than for the other three labs (Figure 3.4). There is noticeable within-lab, between-run variation as can be seen in Figure 3.3.

Theoretically, the true top plateau level should be 100% and the top estimates should center around 100%. In terms of central tendency alone, the top estimates from lab C were quite good, although the apparent tendency of them to increase over time (run numbers reflected chronological order) was a cause for concern. Lab E, which was supposedly the best-performing as explained earlier, did not produce expected results for estradiol. The confidence interval (CI) for the top plateau estimate for estradiol computed for lab E shown in Figure 3.3 did not include 100%, meaning there is statistical evidence that lab E reported top plateau estimates consistently greater than 100% (in Figure 3.4 the CI for estradiol for lab E does include 100%, but the computation of that CI was based on the pretence that we only had a hypothetical 3 runs instead of the actually available 6 runs). The top plateau estimates for estradiol from lab D, on the other hand, had a tendency to be lower than 100% as the CI for that barely included the 100% value. The top estimates from Lab A were highly variable in terms of both within-run and between-run variation.

Let's turn our attention to the estimates of the bottom plateau level, results for which are shown in Figure 3.5 and Figure 3.6. Note that these plots were drawn using a scale directly comparable to that used for the top parameter, i.e., the biologically expected value ± 20%. Except for the norethynodrel results from lab A, most of the time the estimates are close to zero, which biologically is the most credible value. The good accuracy overall for this parameter most likely reflects the fact that computation of y values includes a step to subtract an estimate of non-specific binding, which is measured for each run . There is strong evidence that bottom estimates are often lower than 0%. This could be explained if the non-specific binding value, which is estimated using a

fixed, high concentration of non-radioactive estradiol for each run, tended to be overestimated.[64]

[64] There seems to be no readily available explanation for why the non-specific binding (NSB) value may have been consistently overestimated. Some speculations on the potential cause of the overestimation are made below.

In an ideal assay, we would quantify NSB at each analyte concentration and subtract it from the total binding to get at the specific binding. This is not done in the current protocol, and most likely such NSB measurements would practically be impossible. Instead, we use a single NSB value for each run, which is estimated by measuring the total binding in the presence of excess amount of non-radioactive estradiol and in the absence of a test chemical. This is akin to the general idea of background signal subtraction, but the analogy does not work perfectly since the NSB would change according to the concentration of test chemical.

The single, nominal NSB value thus obtained may tend to be greater than the actual levels of NSB at varying concentration of the test chemical. Theoretically the actual level of NSB increases proportionally with the concentration of free radioactive estradiol in the assay solution. The actual levels of NSB could have been lower than the nominal level if the concentration of free radioactive estradiol present in the liquid phase at varying chemical concentrations had been lower than the level corresponding to the fixed excess level of non-radioactive estradiol used for estimation of the common non-specific binding value. Since we do not know the actual levels of NSB this possibility remains to be a speculation.

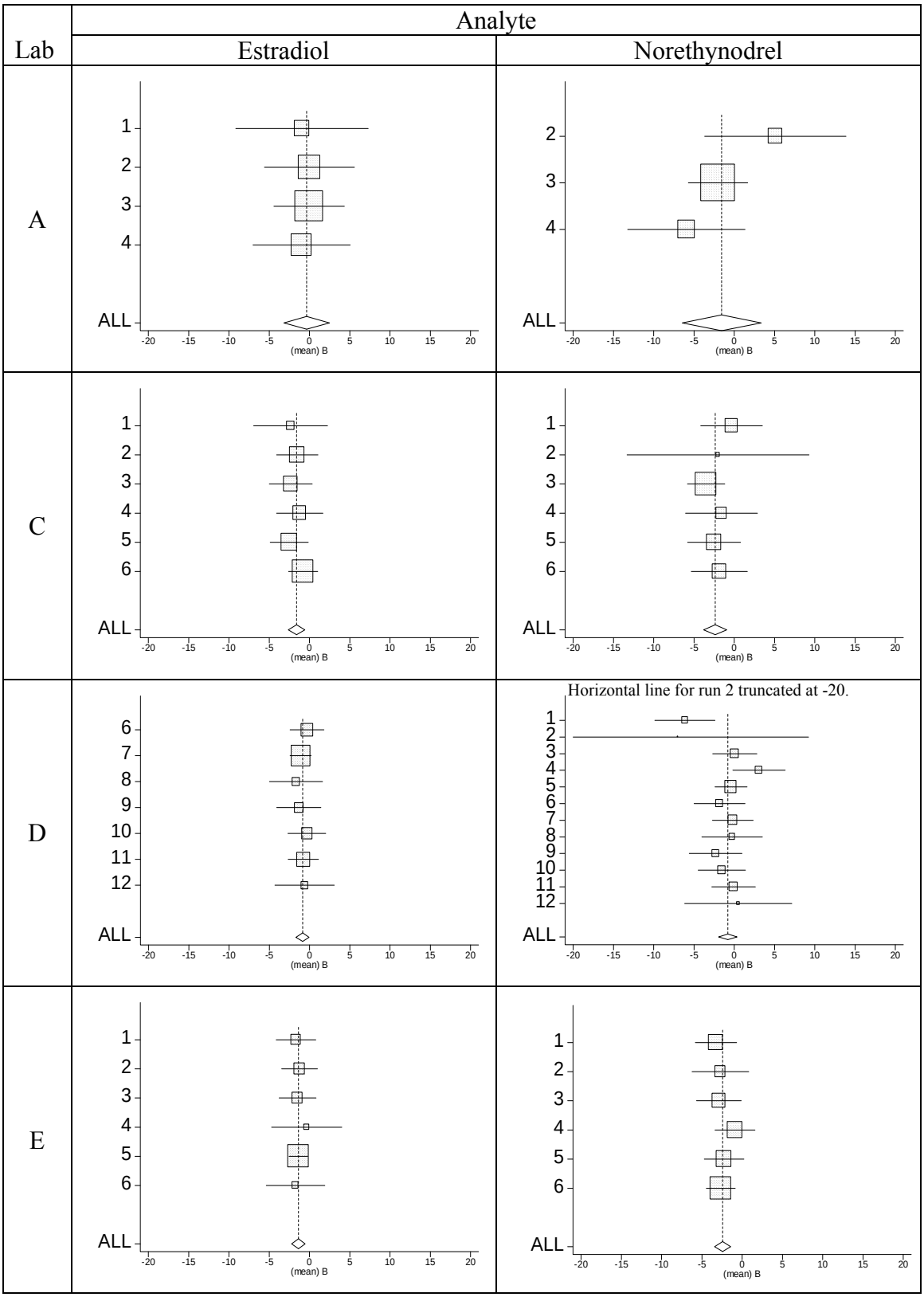

Figure 3.5 Bottom parameter estimates by analyte and laboratory

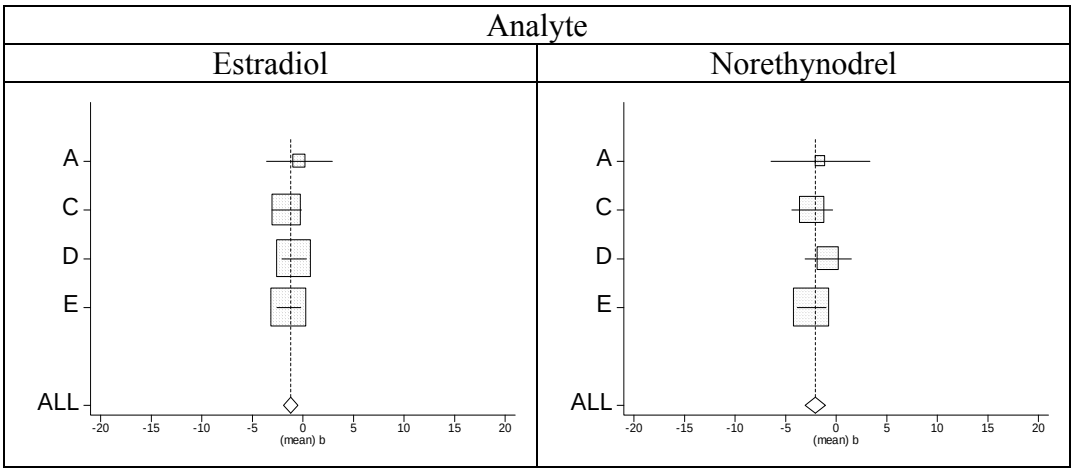

Figure 3.6 Across-lab summary of the estimates of the bottom plateau level

Whatever the mechanism is, the tendency for the laboratories to report negative bottom plateau levels may be an artifact of the protocol rather than an indication of improper performance of the laboratories as a whole.

The results for Hill slope estimates are given in Figure 3.7 and Figure 3.8. Most of the time the 95% CI from a run includes the biologically expected value of -1. Except for the tendency for lab A to produce a Hill slope estimate for estradiol < -1 all labs produced Hill slope estimates near -1.

The pooled across-lab summary of Hill slope was slightly greater than -1, but its 95% CI included -1. Overall, the Hill slope estimates appeared to be credible and deviation from -1, if any, seemed to be due to reasonably small random error.

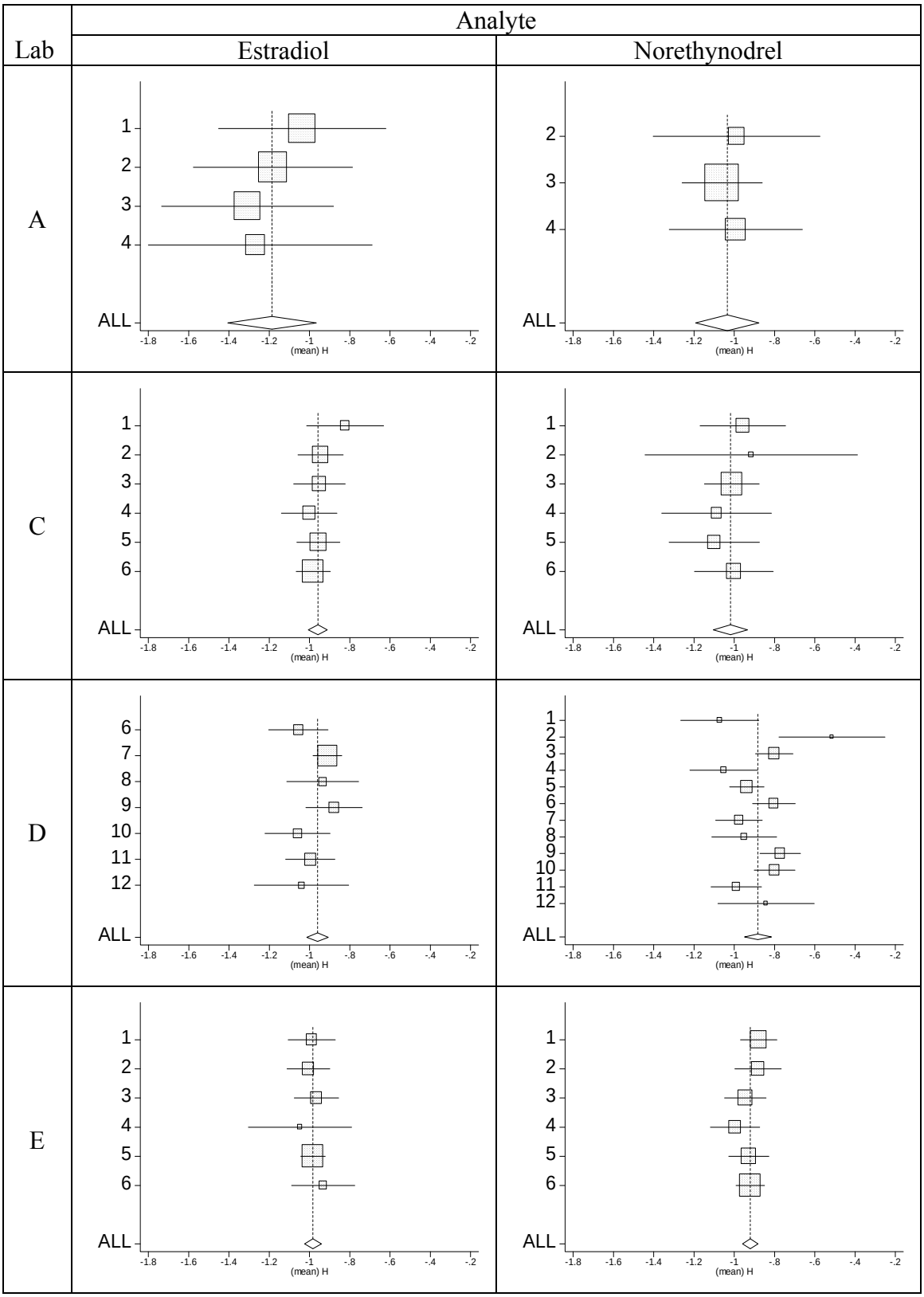

Figure 3.7 Hill slope parameter estimates by analyte and laboratory

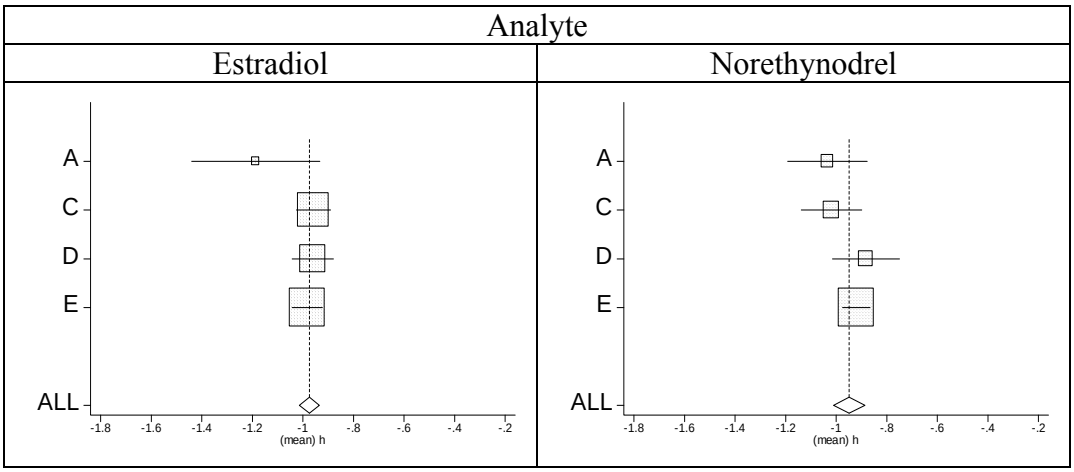

Figure 3.8 Across-lab summary of the estimates of the Hill slope

Results for logRBA for norethynodrel are shown in Figure 3.9 and Figure 3.10. Note that the same scale is used in these figures as is used in Figure 3.1 and Figure 3.2 for $\log IC_{50}$, for easy visual comparison. Computation of logRBA requires a matched pair of $\log IC_{50}$ estimates for estradiol and test chemical, and so there are only 7 logRBA estimates for lab D although there are 12 $\log IC_{50}$ measurements for norethynodrel.

Comparing Figure 3.9 to Figure 3.1, within each lab the between-run scatter of logRBA seems to be smaller than that of $\log IC_{50}$ for labs A, C, and D. Comparing Figure 3.10 to Figure 3.2, the between-lab scatter of logRBA is smaller than that of $\log IC_{50}$. These presumably are the result of within-run correlation between (i.e., parallel shift for) $\log IC_{50}$ of standard and that of norethynodrel. The width of the diamond, i.e., the width of the 95% CI for overall mean, in Figure 3.10 is much narrower than that in the right-hand panel of Figure 3.2, indicating the overall between-lab variation is smaller for logRBA than for $\log IC_{50}$.

The across-lab order of $\log IC_{50}$ values for norethynodrel and that of logRBA for norethynodrel were different. Norethynodrel $\log IC_{50}$s (as well as standard $\log IC_{50}$s) for lab D were the highest among the four labs whereas norethynodrel logRBA for lab D was the lowest. This observation that the order was not preserved between $\log IC_{50}$ and logRBA of the same chemical may be taken as a sign, if not a very convincing one, that the standardization by standard's $\log IC_{50}$ was helpful in reducing between-lab variation. High precision of logRBA for lab E also is noticeable. The concerns over potential qualitative[65] differences between lab E and the remaining labs have been explained in the footnote 63, and they are applicable here.

[65] Dr. Feder suggests that "qualitative(ly)" used in this context be replaced by "systematic(ally)". I prefer to retain "qualitative(ly)". My choice of the adjective "qualitative" was based on its broad meaning. Although lab E seems to be systematically in a quantitative manner from other labs (i.e., different $\log IC_{50}$ values and smaller variation in $\log IC_{50}$), we do not know exactly how the difference came about. The word "systematic" seems to imply we are to some extent certain about the exact nature of the difference in question. What we are certain about the difference is the fact that lab E was the leading lab and improvement for the protocol was largely based on its experience. My view is that such "qualitative" has bring about the systematic difference in the parameter estimates. At any rate, if a reader would like to

If the standard-norethynodrel shift for $logIC_{50}$ is completely parallel, logRBA estimates would show very low or zero between-run and between-lab variation. That was not the case. Nonetheless, as can be seen in comparisons of numerical results shown in Table 3.4 and Table 3.12 to be presented later, overall and intrinsic between-run/lab variabilities were generally lower for logRBA of norethynodrel than for $logIC_{50}$ of norethynodrel (except for overall and intrinsic between-run variabilities for lab E, where these variabilities for norethynodrel $logIC_{50}$ were rather low).

The fact that the confidence interval for logRBA from one of the laboratories excludes the pooled-across-lab-mean is not so much to worry about. Looking at Figure 3.10 from another perspective, the point estimate of logRBA for lab E lies, if barely, within the confidence interval of the pooled across lab mean (i.e., the estimate is inside the edges of the diamond). The within-lab variability of logRBA is expected to decrease as the protocol is refined. With that change, the width of the confidence interval for each lab gets narrower, and the chances that the confidence intervals for some labs exclude the pooled across-lab mean would increase if the intrinsic between-lab variability of logRBA remains unchanged.

What we need to bear in mind is that lab E actually was qualitatively different from other labs[66]. It was the leading lab, which contributed to the improvement of the protocol. Other labs received the improved protocol and started from scratch. In terms of the level of experience with the RUC estrogen receptor binding assay, the laboratories that will take part in subsequent interlaboratory studies and actual implementation of this assay will likely be more like labs A, C, and D than lab E. That implies the performance criteria we developed using data from lab E are not optimal to be applied to novice laboratories.

As we reevaluate the performance criteria in the light of the new data from subsequent interlaboratory study (and update them if necessary), it would be prudent to make efforts to remove the influence of lab E on the update performance criteria.[67] Sensitivity analyses, e.g., comparing criteria derived with and without inclusion of data from lab E, would facilitate such efforts. In deriving the criteria described in this report, sensitivity analyses of this kind were not feasible because of the small number of laboratories from which usable data were available. For instance, we determined that labs D and E were acceptable in terms of precision of $logIC_{50}$, and removal of data from lab E would have left data from a single lab, from which no estimate of between-lab variability could be computed. Without an estimate of between-lab variability, the methods for deriving performance criteria that are described in this report do not work.

---

replace the original wording with "systematic(ally)" it can be done so without altering the originally intended nuances.

[66] Dr. Feder would like to insert a phrase "with respect to IC_{50} but not RBA". It is true in terms of location of $logIC_{50}$ and logRBA, but, as noted earlier in the body of text, lab E also has smaller variation for logRBA. As such, I prefer to keep the original sentence as is.

[67] This advise is expressed from a practical and empirical view point. From a purely scientific view point, laboratories should meet the higher standards of a well-trained lab such as lab E.

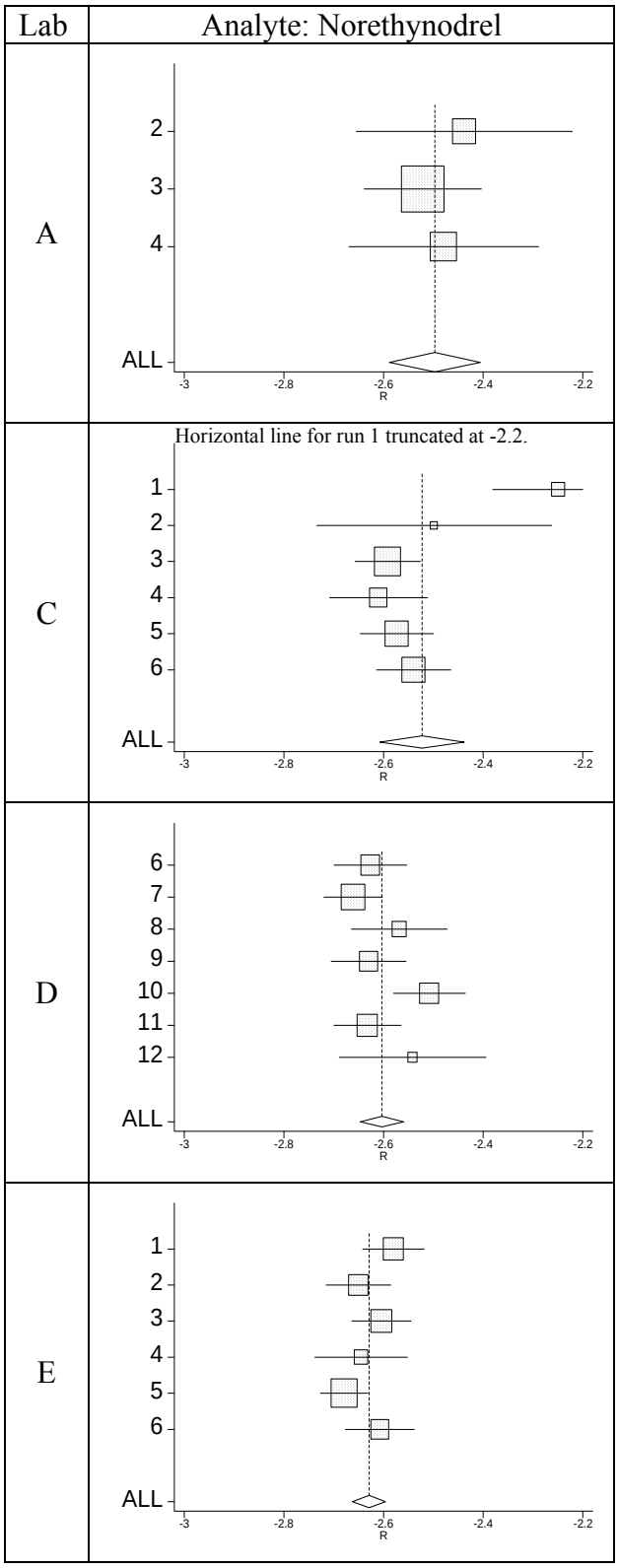

Figure 3.9 logRBA parameter estimates by laboratory

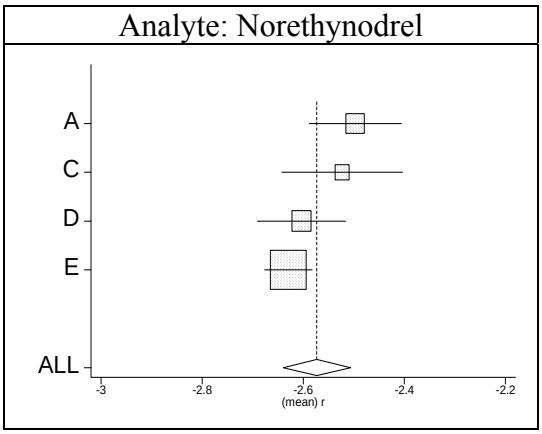

Figure 3.10 Across-lab summary of logRBA parameter estimates

3.4 Within-run, between-run (= within-lab), and between-lab variations

The forest plots are useful in conveying an overall impression of within-unit variation and overall between-unit variation at two levels, run and lab, of hierarchy. They do not, though, help us gain quantitative insights on these. They also are less informative about intrinsic between-unit variation. This section provides a quantitative summary of the various variability measures as well as some examples of how they are computed.

Let us use $\log IC_{50}$ data on estradiol from lab A, which are shown below, as an example.[68] The notation used in "2.2.1 DerSimonian-Laird random effects model" will be used as consistently as possible.

Run (i)	$\hat{\theta}_i$	$SE(\hat{\theta}_i)$ [69]
1	-8.792	0.0769
2	-8.956	0.0510
3	-8.971	0.0425
4	-9.107	0.0564

Fitting the DL random effects model to this data yields the output shown below.

```
Meta-analysis

       |  Pooled      95% CI          Asymptotic      No. of
Method |     Est   Lower   Upper  z_value  p_value   studies
-------+-------------------------------------------------------
Fixed  |  -8.976  -9.028  -8.924 -338.468    0.000      4
Random |  -8.965  -9.070  -8.861 -168.064    0.000
```

[68] Using computation for $\log IC_{50}$ as an example is not the best choice since across-lab comparison of $\log IC_{50}$s is not something that the EDSP regards valid at the time of this writing. This choice merely reflected our initial focus in 2005 on $\log IC_{50}$.

[69] Transcriptional errors for data from runs 2-4 in the draft report have been corrected.

```
Test for heterogeneity: Q= 11.298 on 3 degrees of freedom (p= 0.010)
Moment-based estimate of between studies variance =  0.008
```

For our purposes, we ignore the results based on fitting the "fixed effects model", which are shown on the row labeled "Fixed". Among the statistics of our interest computed from the DL methods, only the overall mean and heterogeneity *p*-value are shown in **bold** in Table 3.2.

Table 3.2 Summary statistics related to within- and between-run variation

Lab	Overall mean $\hat{\theta}_R$	SE(overall mean) $\widehat{SE}(\hat{\theta}_R)$ [70]	Overall between-run SD $\widehat{SD}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$	Within-run SD $SD_{within\text{-}run}(\hat{\theta}_i)$	Intraclass correlation	Between/ within ratio	Heterogeneity *p*-value
A	**-8.965**	0.0533	0.1067	0.0905	0.0565	0.720	1.60	**0.010**

The other statistics are calculated from interim results, which do not appear in the output and need to be extracted in a manner specific to the statistical package in use. The formulae that are used for generating interim results are shown below so that the process can be reproduced irrespective of the statistical package in use.

SE(overall mean), $\widehat{SE}(\hat{\theta}_R)$ = 0.053

$$SE(\hat{\theta}_R) = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}} \quad \text{.....2.17}$$

Overall between-run SD, $\widehat{SD}(\hat{\theta})$ = 0.107 = sqrt(0.053^2*4)

$$SD(\hat{\theta}) = \sqrt{\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}} \quad \text{.....2.18}$$

Intrinsic between-run SD, $\hat{\tau}_{run}$ = 0.091 = sqrt(0.08)

Take the square root of either side of the following.

$$\hat{\tau}^2 = \max\left(0, \frac{\text{Q-(k-1)}}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}\right) \quad \text{.....2.8}$$

[70] The formula should look like $\widehat{SE}(\hat{\theta}_R)$. That is, there should be a hat mark above SD. This is to indicate that the SD is an estimate. It may not be displayed properly depending on the version of the word processor or associated software. The same issue seems to exist for Table 3.2 though Table 3.13.

where

$$w_i = 1/\hat{v}_i \quad \text{.....2.9}$$

$$\hat{\theta}_{\mathrm{F}} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} \quad \text{.....2.10}$$

and $\hat{v}_i$ is the estimated variance of $\hat{\theta}_i$, i.e.,

$$\hat{v}_i = SE(\hat{\theta}_i)^2 \quad \text{.....2.11}$$

Within-run SD, $SD_{\text{within-run}}(\hat{\theta}_i)$ = `0.056 = sqrt(0.053^2*4 - 0.008)`

$$SD_{\text{within-lab}}(\hat{\theta}_{\mathrm{i}}) = \sqrt{\left(\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}\right) - \hat{\tau}^2} \quad \text{.....2.23}$$

Intraclass correlation = `0.849 = 0.008/0.107^2`

$$\text{Intraclass Correlation} = \hat{\tau}^2 / SD(\hat{\theta})^2 \quad \text{.....2.19}$$

Between/within ratio =
(Intrinsic between-run SD)/(within-run SD) = `1.6 =0.091/0.056`
Heterogeneity *p*-value = `0.10` = $\Pr(q > Q = 11.298 |\ \mathrm{q} \sim \chi^2(3))$

$\Pr(q > Q |\ \mathrm{q} \sim \chi^2(k-1))$, i.e., the Q statistic follows the chi-square distribution of degrees of freedom *k-1*, *k* being the number of units, and

$$\mathrm{Q} = \sum_{i=1}^{k} w_i (\hat{\theta}_i - \hat{\theta}_{\mathrm{F}})^2 \quad \text{.....2.9}$$

where

$$w_i = 1/\hat{v}_i \quad \text{.....2.10}$$

$$\hat{\theta}_{\mathrm{F}} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} \quad \text{.....2.11}$$

Similar computation for labs C, D, and E generates a set of corresponding run-specific summary measures. They are summarized in Table 3.3 for estradiol and Table 3.4 for norethynodrel. Once lab-specific summaries are computed, we can further summarize them across laboratories. The across-lab summary so obtained also is included in Table 3.3 and Table 3.4.

Table 3.3 Within- and between-variabilities for logIC$_{50}$: Estradiol[71]

Lab	Overall mean $\hat{\theta}_{\mathrm{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\mathrm{R}})$	Overall between-run SD $\widehat{SD}_{run}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$	Within-run SD $\widehat{SD}_{\text{within-run}}(\hat{\theta}_i)$	Intraclass correlation	Between/ within ratio	Hetero-geneity *p*-value
A	-8.965	0.0533	0.1067	0.0905	0.0565	0.720	1.60	0.01
C	-8.966	0.0328	0.0803	0.0765	0.0242	0.909	3.16	$8*10^{-14}$
D	-9.158	0.0293	0.0774	0.0740	0.0228	0.913	3.25	$6*10^{-17}$
E	-8.905	0.0164	0.0403	0.0331	0.0229	0.676	1.45	0.002
	Overall mean $\hat{\theta}_{\bar{\mathrm{R}}}$[72]	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{\mathrm{R}}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta}_i)$	Intraclass correlation	Between/ within ratio	Hetero-geneity *p*-value
All	-8.997	0.0600	0.1200	0.1112	0.0450	0.859	2.47	$1*10^{-5}$

Table 3.4 Within- and between-variabilities for logIC$_{50}$: Norethynodrel

Lab	Overall mean $\hat{\theta}_{\mathrm{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\mathrm{R}})$	Overall between-run SD $\widehat{SD}_{run}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$	Within-run SD $\widehat{SD}_{\text{within-run}}(\hat{\theta}_i)$	Intraclass correlation	Between/ within ratio	Hetero-geneity *p*-value
A	-6.517	0.0574	0.0994	0.0707	0.0698	0.507	1.01	0.14
C	-6.463	0.0722	0.1768	0.1679	0.0553	0.902	3.04	$8*10^{-20}$
D	-6.583	0.0266	0.0922	0.0826	0.0411	0.802	2.01	$1*10^{-12}$
E	-6.271	0.0108	0.0264	0.0143	0.0221	0.295	0.65	0.21
	Overall mean $\hat{\theta}_{\bar{\mathrm{R}}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{\mathrm{R}}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta}_i)$	Intraclass correlation	Between/ within ratio	Hetero-geneity *p*-value
All	-6.457	0.1026	0.2053	0.1998	0.0473	0.947	4.23	$1*10^{-28}$

The steps taken to compute the across-lab summary are illustrated below using the estradiol logIC$_{50}$ data as an example.

As explained in "2.2.2 Correction for the number of runs performed", correction for the number of runs performed is applied to standard errors computed for each lab (Table 3.5).[73]

[71] After these tables that summarized within- and between-variabilities were completed, it was realized that inclusion of statistics based on the conventional method, e.g., overall between-run SD, overall between-lab SD, and intraclass correlation computed over labs as units of observation, could have been informative.

[72] To distinguish the between-lab overall mean from within-lab, between-run overall mean, the former is designated as $\hat{\theta}_{\bar{\mathrm{R}}}$ rather than $\hat{\theta}_{\mathrm{R}}$.

[73] Please see the addendum on the "number of runs" correction (Section 4.2 Improved correction for the number of runs performed).

Table 3.5 "Number of runs" correction for standard errors

Lab	Lab-specific summaries in Table 3.3		Number of runs performed	$SE(\hat{\theta}_i)$ for 3 runs
	$\hat{\theta}_i$	$SE(\hat{\theta}_i)$		
A	-8.965	0.0533	4	0.0616
C	-8.966	0.0328	6	0.0464
D	-9.158	0.0293	7	0.0447
E	-8.905	0.0164	6	0.0233

e.g., for lab A, $\sqrt{\frac{1}{\left(\sum_{j=1}^{k_i}\frac{1}{SE(\hat{\theta}_{ij})^2+\hat{\tau}_i^{\,2}}\right)\frac{3}{k_i}}}=\sqrt{\frac{1}{\left(\sum_{j=1}^{k_i}\frac{1}{SE(\hat{\theta}_{ij})^2+\hat{\tau}_i^{\,2}}\right)}}\frac{1}{\sqrt{\frac{3}{k_i}}}=\frac{0.0533}{\sqrt{3/4}}=0.0616$

$$SE(\hat{\theta}_{\text{R for lab i}})=\sqrt{\frac{1}{\left(\sum_{j=1}^{k_i}\frac{1}{SE(\hat{\theta}_{ij})^2+\hat{\tau}_i^{\,2}}\right)\frac{3}{k_i}}} \quad \text{.....2.20}$$

Using values of $\hat{\theta}_i$ and "$SE(\hat{\theta}_i)$ for 3 runs" in the thick rectangles of Table 3.5, statistics in the last row of Table 3.3 are computed.

Similar computation for the top, bottom, Hill slope, and logRBA generates the results shown in Table 3.6 through Table 3.12. The numerical results shown in these tables correspond to the graphical across-lab summaries presented in the preceding section "3.3 Run- and lab-specific estimates of the Hill equation parameters".

Table 3.6 Within- and between-variabilities for top parameter: Estradiol

Lab	Overall mean $\hat{\theta}_{\bar{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{R}})$	Overall between-run SD $\widehat{SD}_{run}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$	Within-run SD $\widehat{SD}_{\text{within-run}}(\hat{\theta}_i)$	Intraclass correlation	Between/within ratio	Heterogeneity *p*-value
A	98.74	4.19	8.39	7.77	3.15	0.859	2.47	$5*10^{-5}$
C	100.30	1.49	3.64	3.20	1.73	0.775	1.85	$6*10^{-5}$
D	97.97	1.05	2.77	2.18	1.71	0.619	1.27	0.008
E	102.76	1.16	2.83	2.47	1.38	0.763	1.80	$3*10^{-5}$
	Overall mean $\hat{\theta}_{\bar{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{R}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta}_i)$	Intraclass correlation	Between/within ratio	Heterogeneity *p*-value
All	100.23	1.28	2.55	1.46	2.09	0.326	0.70	0.021

Table 3.7 Within- and between-variabilities for top parameter: Norethynodrel

Lab	Overall mean $\hat{\theta}_{\bar{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{R}})$	Overall between-run SD $\widehat{SD}_{run}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$ [74]	Within-run SD $\widehat{SD}_{\text{within-run}}(\hat{\theta}_i)$	Intraclass correlation	Between/within ratio	Heterogeneity *p*-value
A	94.03	3.28	5.69	5.04	2.63	0.785	1.91	0.007
C	99.94	1.09	2.68	2.11	1.65	0.622	1.28	0.007
D	99.48	1.05	3.63	3.23	1.65	0.794	1.96	$8.7*10^{-14}$
E	98.93	1.07	2.62	2.47	0.86	0.893	2.89	$1*10^{-09}$
	Overall mean $\hat{\theta}_{\bar{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{R}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta}_i)$	Intraclass correlation	Between/within ratio	Heterogeneity *p*-value
All	99.01	0.92	1.84	0.00	1.84	0.000	0.00	0.44

[74] In Table 3.7 through Table 3.13, some estimates of intrinsic between-run(lab) SD, intraclass correlation, and between/within ratio are zero. These estimates of zero generally occurred as a result of truncation for intrinsic between-run(lab) SD as indicated in the following formula.

$$\hat{\tau}^2 = \max\left(0, \frac{\text{Q-(k-1)}}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}\right) \quad \text{.....2.8}$$

Table 3.8 Within- and between-variabilities for bottom parameter: Estradiol

Lab	Overall mean $\hat{\theta}_{\mathrm{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\mathrm{R}})$	Overall between-run SD $\widehat{SD}_{run}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$ [75]	Within-run SD $\widehat{SD}_{\text{within-run}}(\hat{\theta}_i)$	Intraclass correlation	Between/within ratio	Heterogeneity *p*-value
A	-0.35	1.43	2.87	0.00	2.87	0.000	0.00	0.99
C	-1.60	0.52	1.28	0.00	1.28	0.000	0.00	0.88
D	-0.87	0.41	1.08	0.00	1.08	0.000	0.00	0.99
E	-1.38	0.43	1.05	0.00	1.05	0.000	0.00	1.00
	Overall mean $\hat{\theta}_{\bar{\mathrm{R}}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{\mathrm{R}}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta}_i)$	Intraclass correlation	Between/within ratio	Heterogeneity *p*-value
All	-1.21	0.37	0.73	0.00	0.73	0.000	0.00	8.2E-01

Table 3.9 Within- and between-variabilities for bottom parameter: Norethynodrel

Lab	Overall mean $\hat{\theta}_{\mathrm{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\mathrm{R}})$	Overall between-run SD $\widehat{SD}_{run}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$	Within-run SD $\widehat{SD}_{\text{within-run}}(\hat{\theta}_i)$	Intraclass correlation	Between/within ratio	Heterogeneity *p*-value
A	-1.56	2.49	4.31	2.94	3.15	0.465	0.93	0.16
C	-2.37	0.73	1.79	0.00	1.79	0.000	0.00	0.82
D	-0.80	0.59	2.03	1.14	1.68	0.313	0.68	0.12
E	-2.43	0.51	1.25	0.00	1.25	0.000	0.00	0.84
	Overall mean $\hat{\theta}_{\bar{\mathrm{R}}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{\mathrm{R}}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta}_i)$	Intraclass correlation	Between/within ratio	Heterogeneity *p*-value
All	-2.06	0.52	1.03	0.00	1.03	0.000	0.00	0.67

[75] See the footnote on the previous page. $\hat{\tau}$ of zero usually occurs as a result of truncation.

Table 3.10 Within- and between-variabilities for Hill slope parameter: Estradiol

Lab	Overall mean $\hat{\theta}_{\mathrm{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\mathrm{R}})$	Overall between-run SD $\widehat{SD}_{run}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$ [76]	Within-run SD $\widehat{SD}_{\text{within-run}}(\hat{\theta}_{l})$	Intraclass correlation	Between/ within ratio	Heterogeneity *p*-value
A	-1.19	0.112	0.224	0.000	0.224	0.000	0.00	0.83
C	-0.96	0.024	0.059	0.000	0.059	0.000	0.00	0.75
D	-0.96	0.027	0.072	0.027	0.067	0.136	0.40	0.33
E	-0.98	0.112	0.224	0.000	0.224	0.000	0.00	0.97
	Overall mean $\hat{\theta}_{\bar{\mathrm{R}}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{\mathrm{R}}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta}_{l})$	Intraclass correlation	Between/ within ratio	Heterogeneity *p*-value
All	-0.98	0.021	0.052	0.000	0.052	0.000	0.00	0.97

Table 3.11 Within- and between-variabilities for Hill slope parameter: Norethynodrel

Lab	Overall mean $\hat{\theta}_{\mathrm{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\mathrm{R}})$	Overall between-run SD $\widehat{SD}_{run}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$	Within-run SD $\widehat{SD}_{\text{within-run}}(\hat{\theta}_{l})$	Intraclass correlation	Between/ within ratio	Heterogeneity *p*-value
A	-1.03	0.080	0.139	0.000	0.139	0.000	0.00	0.92
C	-1.02	0.043	0.106	0.000	0.106	0.000	0.00	0.94
D	-0.88	0.034	0.117	0.093	0.071	0.629	1.30	$3*10^{-4}$
E	-0.92	0.020	0.049	0.000	0.049	0.000	0.00	0.71
	Overall mean $\hat{\theta}_{\bar{\mathrm{R}}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{\mathrm{R}}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta}_{l})$	Intraclass correlation	Between/ within ratio	Heterogeneity *p*-value
All	-0.95	0.032	0.064	0.035	0.053	0.310	0.67	0.24

[76] $\hat{\tau}$ of zero mostly occurs as a result of truncation.

Table 3.12 Within- and between-variabilities for logRBA parameter: Norethynodrel

Lab	Overall mean $\hat{\theta}_{\mathrm{R}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\mathrm{R}})$	Overall between-run SD $\widehat{SD}_{run}(\hat{\theta})$	Intrinsic between-run SD $\hat{\tau}_{run}$ [77]	Within-run SD $\widehat{SD}_{\text{within-run}}(\hat{\theta})$	Intraclass correlation	Between/ within ratio	Heterogeneity *p*-value
A	-2.50	0.0464	0.0804	0.0000	0.0804	0.000	0.00	0.79
C	-2.52	0.0431	0.1057	0.0899	0.0555	0.724	1.62	0.0002
D	-2.60	0.0222	0.0588	0.0423	0.0409	0.518	1.04	0.04
E	-2.63	0.0170	0.0416	0.0258	0.0326	0.385	0.79	0.14
	Overall mean $\hat{\theta}_{\bar{\mathrm{R}}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{\mathrm{R}}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta})$	Intraclass correlation	Between/ within ratio	Heterogeneity *p*-value
All	-2.57	0.0340	0.0680	0.0527	0.0430	0.601	1.23	0.05

Table 3.13 Within- and between-variabilities for labs C, D, and E that are deemed acceptable

Analyte	Parameter	Overall mean $\hat{\theta}_{\bar{\mathrm{R}}}$	SE(overall mean) $\widehat{SE}(\hat{\theta}_{\bar{\mathrm{R}}})$	Overall between-lab SD $\widehat{SD}_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $\widehat{SD}_{\text{within-lab}}(\hat{\theta})$	Heterogeneity *p*-value
Estradiol	$\log IC_{50}$	-9.006	0.0768	0.1330	0.1271	0.0391	$3*10^{-06}$
	Top	100.33	1.51	2.61	1.92	1.77	0.11
	Bottom	-1.25	0.37	0.65	0.00	0.65	0.73
	Slope	-0.969	0.0199	0.0345	0.0000	0.0345	0.84
Norethynodrel	$\log IC_{50}$	-6.434	0.1199	0.2076	0.1975	0.0639	$4*10^{-08}$
	Top	99.44	0.96	1.66	0.00	1.66	0.90
	Bottom	-2.08	0.53	0.91	0.00	0.91	0.47
	Slope	-0.935	0.0320	0.0554	0.0305	0.0463	0.26
	logRBA	-2.606	0.0260	0.0451	0.0243	0.0379	0.26

[77] $\hat{\tau}$ of zero mostly occurs as a result of truncation.

3.5 Deriving accuracy criteria[78]

3.5.1 $logIC_{50}$

An alternative graphical summary of interlaboratory data is given in Figure 3.11 for the estimated $logIC_{50}$ of estradiol. The distribution of $logIC_{50}$ averaged for 3 runs that would be reported from each lab is plotted as a Normal curve.

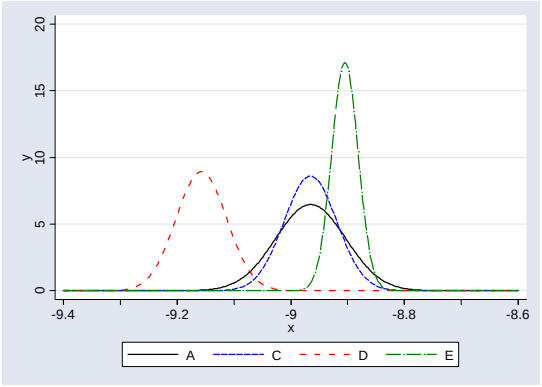

Figure 3.11 Distribution of lab-specific $logIC_{50}$(estradiol) estimates by laboratory

The first step in developing accuracy performance criteria for $logIC_{50}$ is to make an informal decision regarding which of the participating laboratories were deemed acceptable in terms of their performance. A visual summary such as Figure 3.11 along with other plots presented so far would be informative in making this decision.

Lab A was deemed unacceptable because of low data quality. Lab A's low data quality was strikingly evident in the analyses performed before one of the norethynodrel runs from lab A, which was obviously problematic, was excluded from analyses. The distribution curve for lab A in Figure 3.11 was estimated using the data after the exclusion, but lab A still has the most "flat" curve, indicating $logIC_{50}$ estimates from lab A were less precise than those from other labs. In fact, Table 3.3 shows that lab A had the highest overall between-run SD, intrinsic between-run SD, and within-run SD. Its relatively high within-run SD for $logIC_{50}$ seems to be due to high within-run variability in y given x as summarized later. It was felt that a competent laboratory should be able to, after some reasonable attempt to improve data quality, perform better than lab A. We knew that multiple laboratories, i.e., labs C and D, were able to do so.

The next step is to estimate the distribution of $logIC_{50}$s that would be expected from a laboratory drawn from the universe of laboratories that are like labs C, D, and E in terms of their ability to measure $logIC_{50}$ accurately and precisely.[79] The distribution has the

[78] As noted earlier, accuracy criteria for $logIC_{50}$ were derived, but the ESDP has made a decision to not impose them because of the variation in protein concentration in the RUC preparation, which renders the concept of accuracy for $logIC_{50}$ questionable. Nonetheless, the EDSP's effort in 2005 for developing a statistical method for the analyses of data from receptor binding assays focused initially on $logIC_{50}$, and accuracy criteria for $logIC_{50}$ was one of the first criteria that were eventually developed. During the development of $logIC_{50}$-related methods, many relevant issues were realized, investigated, and resolved. It makes easier to address these issues in this report if the historical context is maintained to some extent.

[79] As mentioned earlier, the concept of accuracy is ill-defined for $logIC_{50}$ since the protein concentration of RUC preparation, which is difficult to standardize across laboratories or batches, is a strong determinant of $logIC_{50}$. In our general scheme of deriving accuracy performance criteria for a parameter, the target value of a parameter, which is taken to be "accurate", is estimated using the pooled mean of the estimates of the parameter from labs that are judged to have acceptable data quality. Because of the protein concentration issue, though, the pooled mean of $logIC_{50}$ does not truly have the interpretation of the accurate $logIC_{50}$ value. The method for deriving the accuracy criteria for $logIC_{50}$, is included as an example and what referred to as "accuracy" in section 3.5.1 is nominal accuracy. Its use as an example includes pointing out

mean $\hat{\theta}_{\bar{R}}$ = -9.006 and spread $SD_{lab}(\hat{\theta})$ = 0.1330 as shown in Table 3.13. Based on these parameter estimates, an interval, to which a future realized value of $\log IC_{50}$ will fall at a specified probability, can be computed. As shown earlier in "2.5.1 $\log IC_{50}$, Top, Bottom, Hill slope", such an interval generally called a "prediction interval" is given as

$$\hat{\theta}_R \pm t_{1-\alpha/2,\,\nu} * \sqrt{1+1/k} * SD(\hat{\theta}) \quad \text{.....2.40}$$

Using this, an 80% prediction interval is derived as follows.

-9.006 ± 1.82*sqrt(1+1/3)*0.1330

The estimated distribution of future $\log IC_{50}$s of estradiol from an acceptable lab and the 80% prediction interval are shown in Figure 3.12 along with the overall mean of the distribution of $\log IC_{50}$s for labs similar to labs C, D, and E.

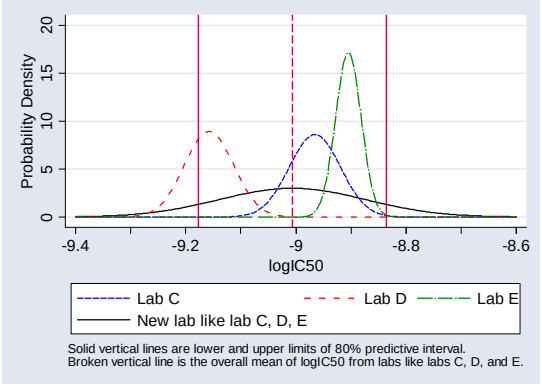

Figure 3.12 Distributions of lab-specific $\log IC_{50}$ for estradiol and prediction interval[80]

The probability associated with the prediction interval, 80% in this case, was chosen based on informal judgment. We do not want to set this probability too high since that would result in accepting values so far away from the middle of the distribution. At the same time, we don't want to set this probability too low, which would make the acceptable ranges too narrow. A few factors would help choose the probability what makes practical sense.

For instance, from Figure 3.12 we can see that if labs C and E were to produce $\log IC_{50}$ values repeatedly, they would fall in the prediction interval most of the time. Note, though, that the curve for lab D indicates that a $\log IC_{50}$ reported by lab D would be accepted only at a probability slightly more than 50% despite the fact that lab D was deemed acceptable to begin with. This could have constituted a basis for increasing the probability to, say, 90%.[81]

This kind of assessment of acceptance rate would help gauge the prospect of identifying acceptable laboratories among any laboratories that might apply for qualification. In the above example, there are three novice labs taking part in the interlaboratory study (i.e., labs A, C, and D), and among them the expected number of acceptable laboratories was about 1.5 (lab A had near 100% acceptance rate, lab C had about 100% acceptance rate,

---

potential issues that may arise in the future and how we might deal with them. It also served as an example of real utility for the assay in which the protein concentration could be better controlled.

[80] The limits shown in this plot are incorrect.

[81] Dr. Feder objects to this practice of "tweaking" probability level. In the light of the need to balance sensitivity and specificity of performance criteria (see Aoki, 2007b for some discussion on this), I believe there is certain justification for changing the probability level.

lab D had about 50% and so expectation is 1*1+1*1+1*0.5 = 2.5). This would give us an estimate of about 83% acceptance rate (2.5/3*100) for all novice lab that would take part in the future qualification process assuming they are like labs A, C, and D. In these calculations we ignored the existence of lab B, which was disqualified at an early stage. If we include lab B in the calculation, the acceptance rate for any novice laboratories would decrease accordingly.

The criteria for norethynodrel were established following similar steps. Figure 3.13 shows the distribution of $\log IC_{50}$ estimates for norethynodrel. Although the distribution for lab A was not worse than others, data from only labs C, D, and E were used, to make the analysis consistent with that for estradiol. Figure 3.14 shows the prediction interval and distribution curves. The distributions for labs D and E have virtually no overlap.

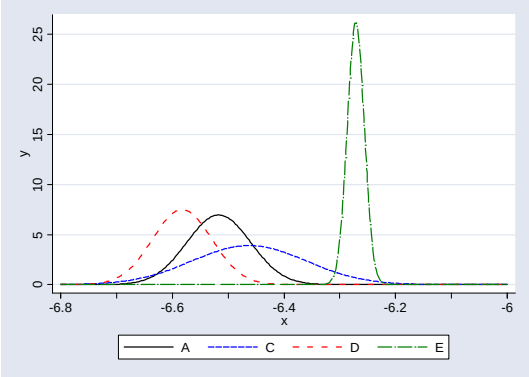

Figure 3.13 Distributions of lab-specific $\log IC_{50}$ estimates for norethynodrel by laboratory

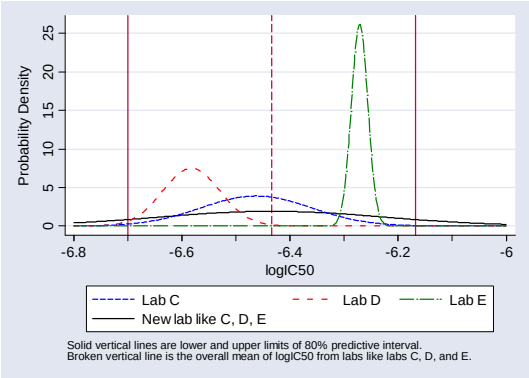

Figure 3.14 Distributions of lab-specific $\log IC_{50}$ estimates for norethynodrel and prediction interval[82]

[82] The limits shown in this plot are incorrect.

3.5.2 Top

Distributions of the estimates of the top plateau for estradiol are plotted in Figure 3.15. The distribution of the estimate from lab A is much flatter than the distributions from other labs. Lab A was deemed unacceptable, and it was decided that the criteria for the top plateau would be derived using the data from labs C, D, and E.

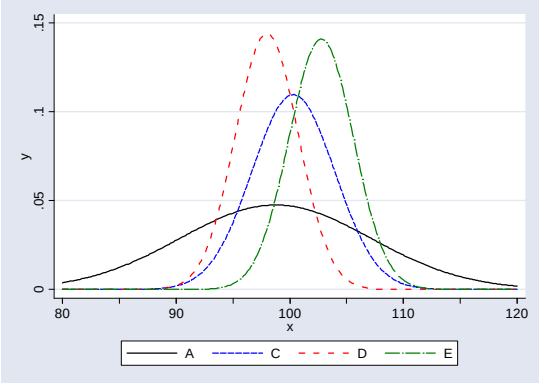

Figure 3.15 Distributions of the estimate of the top plateau for estradiol, by laboratory

The prediction interval for a parameter may be derived either for a lab-specific summary of the parameter based on data from 3 runs, or for a run-specific summary of the parameter based on data from a single run.

For the three Hill equation parameters that have to do with shape, not horizontal location, of the binding curve (i.e., top, bottom, Hill slope), an approach to derive criteria for a run-specific estimate was taken. For each of these parameters, a common biologically plausible level exists for any chemical that interacts with the estrogen receptor through a typical one-site competitive binding mechanism, no matter how strong or weak the interaction is. For any individual run a competent laboratory would generate data from which estimates of these parameters close to the biologically plausible level should be obtained. Because of this increased certainty, it was felt justifiable to set up criteria to which estimates of these parameters from each run of estradiol and norethynodrel are compared. (Estradiol and norethynodrel are typical one-site competitive binders.)

For an estimate from each run, the prediction interval formula is modified slightly as follows to account for the fact that the estimate is run-specific, not a pooled estimate based on data from 3 runs.

$$\hat{\theta}_R \pm t_{1-\alpha/2,\,\nu} * \sqrt{1+1/k} * \widehat{SD}_{lab}(\hat{\theta})\sqrt{3}$$

as opposed to

$$\hat{\theta}_R \pm t_{1-\alpha/2,\,\nu} * \sqrt{1+1/k} * \widehat{SD}_{lab}(\hat{\theta})$$

This is because the $\widehat{SD}_{lab}(\hat{\theta})$ was a lab-specific summary estimate based on 3 runs, not a run-specific summary.

Using the between-lab summary information in Table 3.13, the 95% prediction interval for the top estimate from an estradiol run is computed as

100.33 ± 4.08*sqrt(1+1/3)*2.61*sqrt(3) = (79.0, 121.63)[83]

The probability of 95%, rather than 80% that was used for a prediction interval for lab-specific logIC_{50} estimate from 3 runs, was chosen based on the tentatively chosen requirement that, for data from a laboratory to be accepted, an estimate of top plateau parameter from a run needs to be accepted three times for three consecutive runs. This issue of the predictive probability will be revisited in "3.5.5 logRBA". The interval is shown in Figure 3.16 along with distribution curves.

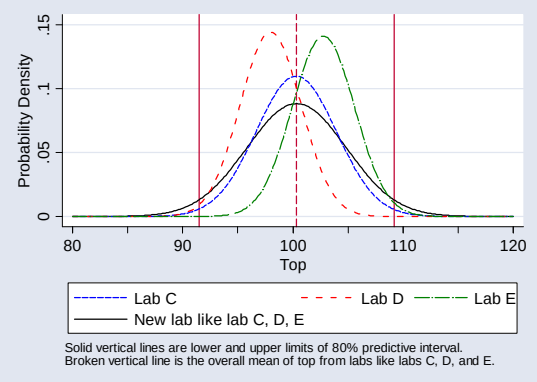

Figure 3.16 Distributions of estimates of the top plateau for estradiol, and the 95% prediction interval[84]

Figure 3.17 shows distributions of the estimates of the top plateau for norethynodrel. Lab A again sets itself apart from the rest. The distribution curves for labs C, D, and E are closer to each other and as a result the 95% prediction interval for norethynodrel, shown in Figure 3.18, is narrower than that for estradiol.

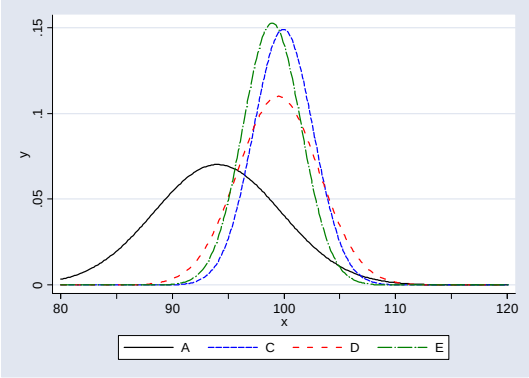

Figure 3.17 Distributions of estimates of the top plateau for norethynodrel, by laboratory

The lower and upper limits for the estimates of the top plateau are presented in Figure 3.18 along with those for bottom and slope parameters, which are established in a similar manner.

[83] The *t* degree of freedom used and the interval computed are copied exactly from Dr. Feder's suggestion on page 108 of his hand-written comments (Feder, 2007b).
[84] The limits shown in this plot are incorrect.

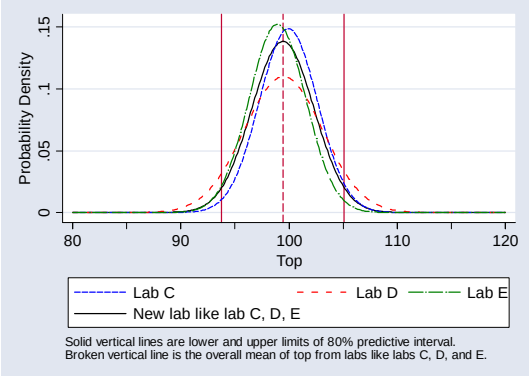

Figure 3.18 Distributions of estimates of the top plateau for norethynodrel, and the 95% prediction interval[85]

3.5.3 Bottom

The distributions for estimates of the bottom plateau for estradiol are close together for labs C, D, and E, and Lab A's estimates are shifted upwards and are more variable (Figure 3.19). Again, the 95% prediction interval was set using data from labs C, D, and E (Figure 3.20).

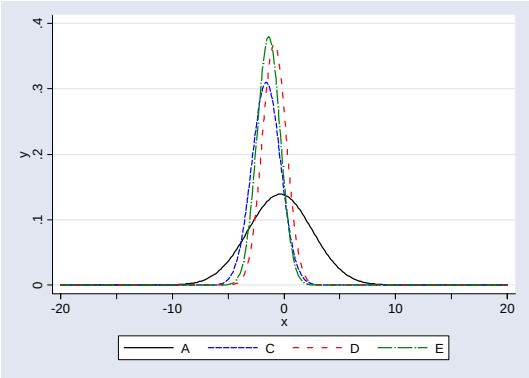

Figure 3.19 Distributions of estimates of the bottom plateau for estradiol by laboratory

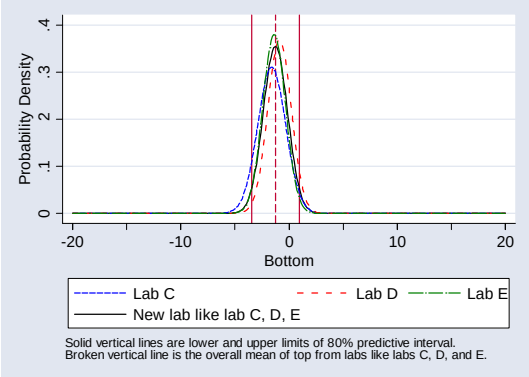

Figure 3.20 Distributions of estimates of the bottom plateau for estradiol, and the 95% prediction interval[86]

[85] The limits shown in this plot are incorrect.
[86] The limits shown in this plot are incorrect.

The distributions for estimates of the bottom plateau for norethynodrel are close together for labs C, and E. Lab D's estimates are shifted upwards with slightly wider spread. Lab A's estimates again are more variable (Figure 3.21). Again, the prediction interval was set using data from labs C, D, and E (Figure 3.22).

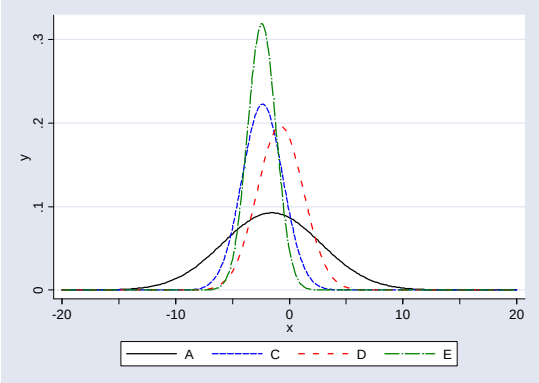

Figure 3.21 Distributions of estimates of the bottom plateau for norethynodrel, by laboratory

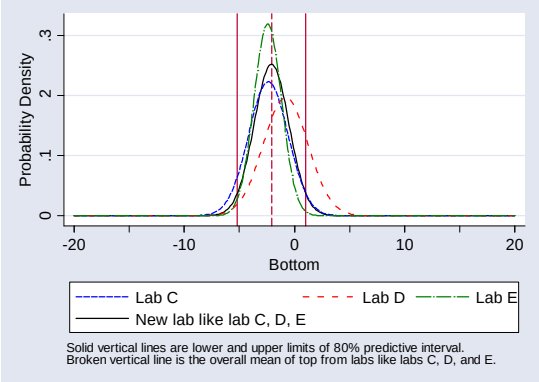

Figure 3.22 Distributions of estimates of the bottom plateau for norethynodrel, and the 95% prediction interval[87]

It is notable that we can see from Figure 3.22 that the bottom plateau estimates from lab D will be accepted only 80% although we included lab D in the set of acceptable laboratories whose data were used to derive the prediction interval.[88] If we would like to ensure that bottom parameter estimates from laboratories exactly like lab D would have a better chance to be accepted, we can increase the predictive probability from 95% to, e.g., 99% (this change would increase the probability of acceptance of the estimates from lab D to 91%). Such a change would cause increased chance of acceptance for estimates from laboratories exactly like lab A, which we deemed unfit, from 53% for the predictable probability of 95% to 65% for the predictable probability of 99%. It is an administrative decision to set the probability level for a prediction interval. Such a

[87] The limits shown in this plot are incorrect.
[88] The discussion in this paragraph is not valid since it is based on incorrectly computed prediction interval.

decision is a balancing act between ensuring better performance of laboratories and potential difficulty in finding such competent laboratories.

As an aid to get a better understanding on the impact of different predictive probability choices, tables similar to Table 3.14 shown below may be produced.

Table 3.14 Acceptance rates of estimates of bottom plateau parameter for two levels of the probability for the prediction interval

	Probability for the prediction interval	
Laboratory	95%	99%
A	53%	65%
C	91%	98%
D	80%	91%
E	98%	100%

3.5.4 Hill slope

Distributions of Hill slope estimates for estradiol and norethynodrel are shown in Figure 3.23 and Figure 3.25, respectively. A large downward shift and higher variability were notable for the estimates for estradiol from lab A. No such distinctive anomaly was seen for the estimates for norethynodrel from lab A, but, because of the overall poor quality of data quality from lab A, only data from labs C, D, and E again were used to derive the prediction intervals. Figure 3.24 and Figure 3.26 show the 95% prediction intervals derived as well as distribution curves.

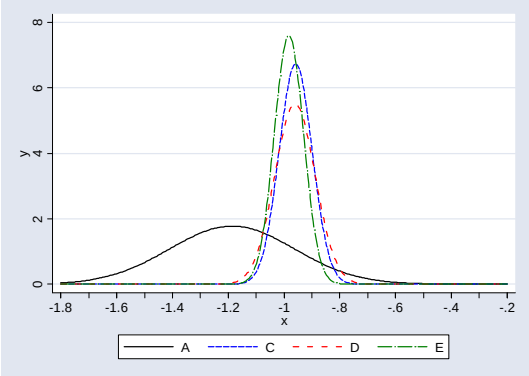

Figure 3.23 Distributions of Hill slope estimates for estradiol, by laboratory

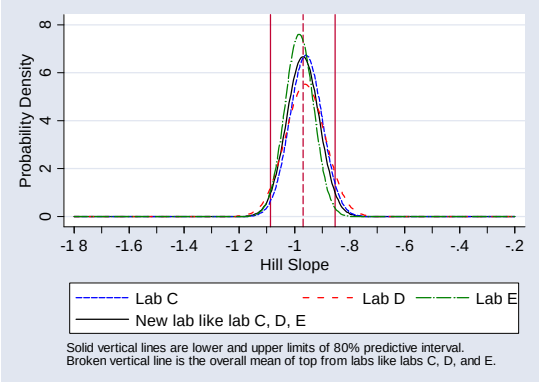

Figure 3.24 Distributions of Hill slope estimates for estradiol, and the 95% prediction interval[89]

[89] The limits shown in this plot are incorrect.

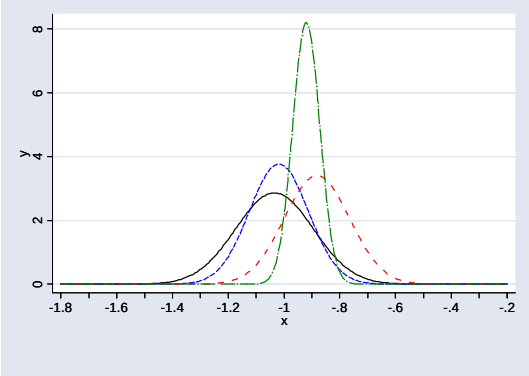

Figure 3.25 Distributions of Hill slope estimates for norethynodrel, by laboratory

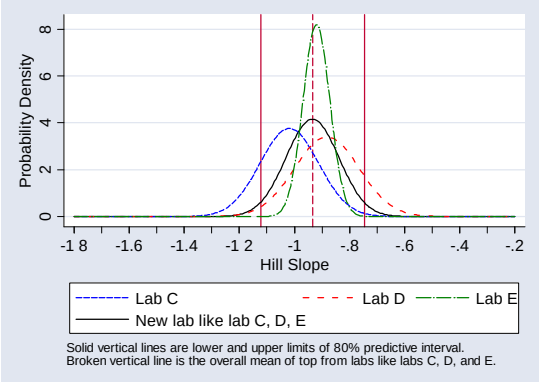

Figure 3.26 Distributions of Hill slope estimates for norethynodrel, and the 95% prediction interval[90]

In Figure 3.26, relatively low rates of acceptance for the estimates from labs C and D (83% and 86%, respectively) may be noted. A similar situation has been noted in Figure 3.22 for lab D, and a potential remedy and its side effects were discussed.[91]

Performance criteria derived for top plateau, bottom plateau, and slope parameters from a single run are summarized in Table 3.15. The number was rounded "outwards" to the nearest number in the unit of 0.1. It may be worth considering increasing upper limits for the bottom plateau parameter because of biological reasons. Both for estradiol and norethynodrel, they seem to be too close to 0.

[90] The limits shown in this plot are incorrect.
[91] The quantitative points in this paragraph is invalid since they are based on incorrectly computed intervals.

Table 3.15 Performance criteria (lower and upper limits of 95% prediction intervals) for top, bottom, slope parameters from a single run[92]

Parameter	Unit	Estradiol		Norethynodrel	
		Lower limit	Upper limit	Lower limit	Upper limit
Top plateau level	[%binding]	~~91.5~~	~~109.2~~	~~93.8~~	~~105.1~~
Bottom plateau level	[%binding]	~~4.0~~	~~1.0~~	~~5.0~~	~~1.0~~
Hill Slope	[(log(M)$^{-1}$]	~~1.1~~	~~0.8~~	~~1.1~~	~~0.7~~

3.5.5 logRBA

The distributions of run-specific logRBA for norethynodrel are shown in Figure 3.27 and included again in Figure 3.28 along with the prediction interval derived for future run-specific logRBA estimates.

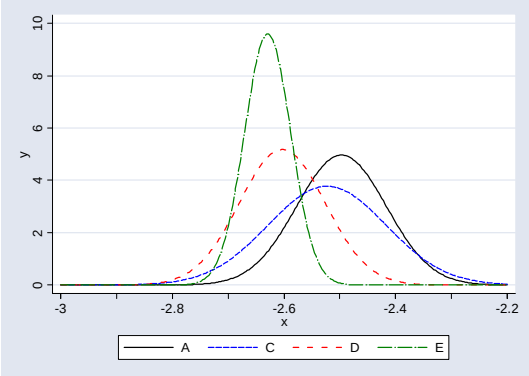

Figure 3.27 Distributions of run-specific logRBA estimates for norethynodrel, by laboratory

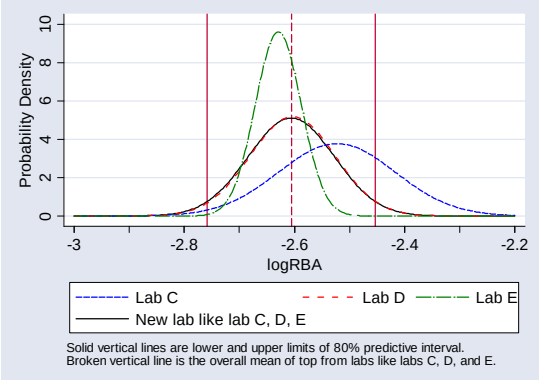

Figure 3.28 Distributions of run-specific logRBA estimates for norethynodrel, and the 95% prediction interval[93]

This prediction interval has not been presented to EDSP for a detailed discussion of its appropriateness. A potential problem may be appreciated in Figure 3.28, which shows

[92] The limits shown in this table are incorrect. The numbers are shown as "strikethrough" to indicated that.
[93] The limits shown in this plot are incorrect.

that logRBA estimates from lab C would be rejected quite often. A simple calculation reveals that the probability of the acceptance is only 73%. Before implementing the criterion for logRBA, its appropriateness should be carefully considered.[94]

A logRBA prediction interval would be used to determine whether a run-specific logRBA is acceptable or not. Instead of setting a prediction interval for run-specific estimates, we can set a prediction interval for lab-specific estimates, i.e., a summary of 3 runs for a new lab. Derivation of such an interval has been shown for $\log IC_{50}$ (Section 3.5.1). For logRBA, imposing a criterion on a lab-specific estimate makes better sense because, unlike top, bottom, and slope parameters for which there is high certainty that their estimates for each *run* are close to biologically plausible values, we would like to make a judgment as to whether a *lab* can produce a logRBA estimate close to the target value.

The distributions and potential criteria for lab-specific logRBAs are shown in Figures 3.29 and 3.30 so that they can be easily contrasted to the run-specific logRBA distributions and criteria that are shown in Figures 3.27 and 3.28.

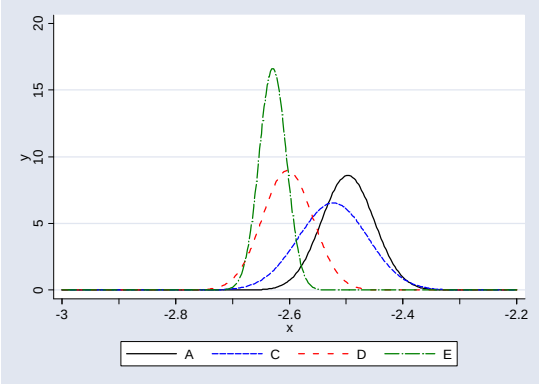

Figure 3.29 Distributions of lab-specific logRBA estimates for norethynodrel, by laboratory

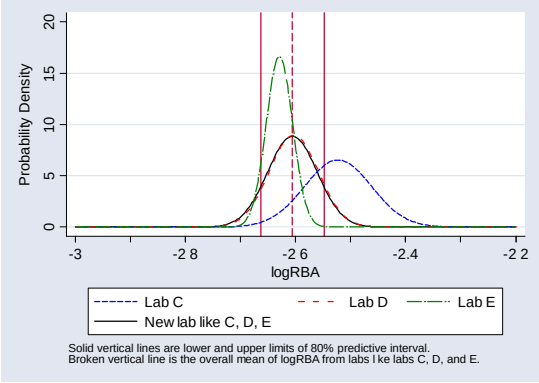

Figure 3.30 Distributions of lab-specific logRBA estimates for norethynodrel, and 80% prediction interval[95]

[94] Because of incorrectly-computed interval limit values, the quantitative points in this and several following paragraphs are invalid.
[95] The limits shown in this plot are incorrect.

Note that for run-specific logRBA estimates a 95% prediction interval was constructed, and for lab-specific logRBA estimates an 80% prediction interval was constructed. A higher coverage probability is used for the run-specific prediction interval since, if we were to require the criterion to be met for three consecutive runs, the overall probability coverage for a set of 3 runs taken together gets smaller ($0.95^3 = 0.853$). An 80% prediction interval for a run-specific logRBA would result in an overall coverage that is too low ($0.8^3 = 0.51$) again if the criterion were to be met for three consecutive runs. In other words, if a run-specific criterion were set so that only 80% of values from a new lab like Labs C, D, and E were accepted, only 51% of such labs would produce 3 consecutive runs that are each acceptable.[96]

A modification to the requirement that 3 consecutive runs meet the criterion would result in a change in overall coverage probability. In order to achieve a desired overall coverage probability, the coverage probability used in deriving the criterion for estimates from a single run could be adjusted.

For instance, we may require that at least 3 out of a total of 5 runs meet the criterion. In this case, the probability for estimates from 3 or more runs falling within the 95% prediction interval is the binomial probability of 3 or more successes in 5 trials when the probability of a success on a single trial is 0.95, which is

$_5C_3$*0.95^3*0.05^2 + $_5C_4$*0.95^4*0.05^1 + $_5C_4$*0.95^5 = 0.9988

If we use the "3 out of 5" requirement but change the probability coverage for a single run from 0.95 to 0.8, this changes to

$_5C_3$*0.8^3*0.05^2 + $_5C_4$*0. 8^4*0.05^1 + $_5C_4$*0. 8^5 = 0.9421

If we want the overall coverage of (X*100)% for the "3 out of 5" requirement, the desired coverage probability for a single run of (P*100)%, we can solve the following equation to obtain P.

$_5C_3$*P^3*(1-P)^2 + $_5C_4$*P^4*(1-P)^1 + $_5C_4$*P^5 = X

An easy way to find P given X is to use a binomial function available in a spreadsheet application such as Excel or statistical package. One could keep trying different values of P as an input to the binomial function of an appropriate form until a desired X is obtained. Such was tried for X = 0.8 and it was found that the acceptance probability of 67.4% for estimates in a single run would achieve an overall coverage probability of 80%.

This calculation gets further complicated if we consider the requirement of meeting criteria for multiple endpoint, e.g., in our case simultaneously meeting the criteria for top plateau, bottom plateau, Hill slope, and logRBA. Ignoring the correlations among estimates of different parameters, a very simple approximation is possible. For instance, the above-mentioned "3 out of 5" requirements with a success probability of 0.95 for a

[96] Dr. Feder recorded many comments regarding the rest of this section. Please see Feder (2007b).

single trial for a single parameter, the overall probability of meeting it simultaneously for 4 parameters would be 0.9988^4 = 0.9952. Changing the success probability to 0.8 would result in the overall probability of 0.9421^4 = 0.7877.

In Figure 3.28 and Figure 3.30, relatively low acceptance for lab C, which was one of the labs deemed acceptable, is noticeable. An observation like this, which also was noted for lab D (Figure 3.22) albeit in a less pronounced manner, is somewhat disturbing. It is not clear how often these occur and whether we need to consider some remedy. Most likely we are hitting the limit for the capacity of the DL random effects model, which is an asymptotic model. Our combined distribution for acceptable labs is based on at most 3 labs so the asymptotic model, which uses a large number approximation, may not match the data well. In order to avoid the consequence of an excessively strict criterion that we might set inadvertently (i.e., rejecting too many laboratories and accepting too few to the extent that the testing program becomes infeasible), it may be advisable to make the criteria somewhat lenient.[97]

3.5.6 Impact of underestimation of SE(lab-specific logIC_{50}) and SE(lab-specific logRBA)[98]

In computing overall across-lab summaries of logIC_{50} and logRBA presented so far, the analytical estimates of standard errors obtained using the DL method were used. These analytical standard errors have been found to underestimate the true level of variability as mentioned in "2.5.1.1 Analytical vs. simulation-based SE(lab-specific logIC_{50})".[99] (Whether the analytical standard error for each run was estimated without bias has not been investigated. If there is a bias of that kind, its impact should be reflected in the analytical standard error for each lab.)

The DL method uses the inverse of estimated within-run variance as a weight, and the underestimation of within-lab variance can result in overestimation of intrinsic between-lab variance $\hat{\tau}^2_{lab}$ since the intrinsic between-lab variance is computed, conceptually speaking, by subtracting the contribution of within-lab variation from the apparent observed variation. The effect of underestimating within-lab variance on the overall

[97] This is likely to be unnecessary if a correct method for deriving a prediction interval had been used as Dr. Feder points out.

[98] Since a decision has been made to not impose accuracy criteria for logIC_{50}, the discussion regarding logIC_{50} in this section has limited relevance. The corresponding discussion on logRBA, however, remains relevant as long as the EDSP is concerned about the accuracy and precision of logRBA estimates. There was a limited amount of data for the analysis of precision of logRBA alone. As explained later, it was discovered the degree of underestimation was similar for logIC_{50} and logRBA and the data for both were combined to obtain a better estimate of magnitude of underestimation.

[99] Dr. Feder suggests the use of maximum likelihood ANOVA in place of DL method. This suggestion seems to be based on the proposition that ML ANOVA generate more valid analytical standard error. Such a proposition does not seem to be supported by the results of the simulation study (Feder and Ma, 2005). As can be seen in Tables 21, 22, 24, 25, 27, and 28 of Feder and Ma (2005), for logIC_{50} for estradiol and norethynodrel, underestimation of standard errors were sometimes greater for ML ANOVA than for DL method for some parameters and laboratories. In some instances, standard errors were overestimated by ML ANOVA. In terms of logRBA, they summarized their finding at these two "procedures performed about the same" (on page xi of Feder and Ma, 2005). As such, the simulation results did not show clear advantage of ML ANOVA over DL.

between-lab variance estimate is difficult to predict because the overall between-lab variance is computed by combining the (underestimated) within-lab variance and the (overestimated) intrinsic between-lab variance. It could be argued that in principle it would be desirable to correct for the underestimation of within-lab variability. No such corrections, however, were made for the summary results presented so far.

The discussion below describes an attempt to correct for average underestimation of standard error of the lab-specific $logIC_{50}$ at the between-run, within-lab level.[100] The analysis revealed that the correction made little difference in the estimates of overall between-lab variability. Based on this finding it was concluded that the correction was unnecessary.[101]

As mentioned earlier, the results of simulation studies performed by Battelle (Feder and Ma, 2005) were used for this comparison. The first step to take for this comparison is to extract necessary information, i.e., average analytical standard error of SE(lab-specific $logIC_{50}$) and simulation-based standard error of SE(lab-specific $logIC_{50}$) are extracted from the report for each combination of laboratory and test chemical. How such information was extracted is shown below using the combination of lab C and norethynodrel as an example.

Table 3.16 Excerpt of "Table 22" for Laboratory C, Norethynodrel by Feder and Ma (2005)[102]

Param-eter		Estimate DL	Average StdErr DL	Run-to-Run Vrnce DL	
B		-2.390 (0.5332)	0.862 (0.2164)	0.063 (0.2996)	
T	Omitted	99.944 (1.5892)	1.410 (0.6718)	5.665 (6.8353)	Omitted
U		-6.466 (**0.1253**)	**0.103** (0.0552)	0.040 (0.0441)	
H		-1.012 (0.0360)	0.050 (0.0141)	0.001 (0.0028)	

0.103 = Average analytical standard error(lab-specific $logIC_{50}$)
0.1253 = Simulation-based standard error(lab-specific $logIC_{50}$)

[100] Dr. Feder repeats his suggestion regarding the use of maximum likelihood ANOVA in place of DL method. As discussed in the footnote 99, this does not seem to be a clear advantage of this alternative method.
[101] The initial attempt to make this comparison in 2005 unfortunately included some methodological flaws. Using a correct method the comparison was performed again as this report was prepared. It has turned out that the methodological flaws did not affect the numerical results much. The comparison results based on the correct method is shown below followed by a brief description of the incorrect method that should have been avoided.
[102] Feder and Ma (2005) used symbols B, T, U, and H to designate bottom plateau, top plateau, $logIC_{50}$, and Hill slope parameters, respectively.

The simulation-based standard error(lab-specific logIC_{50}) represents the true level of random variation that the analytical standard error should center around. For this combination, the analytical standard error on average was smaller than the simulated-based standard error, and the same is the case for all of the other lab-chemical combinations as summarized in Table 3.17.

Table 3.17 Comparison of analytical standard error and simulation-based standard error for logIC_{50}

	Estradiol		Norethynodrel	
	Analytical SE	Simulation-based SE	Analytical SE	Simulation-based SE
C	0.0489	0.0611	0.1034	0.1253
D	0.0366	0.0457	0.0446	0.0571
E	0.0177	0.0226	0.0171	0.0180

Corresponding results for logRBA are shown in Table 3.18.

Table 3.18 Comparison of analytical standard error and simulation-based standard error for norethynodrel logRBA

	Analytical SE	Simulation-based SE
C	0.1146	0.1413
D	0.0580	0.0758
E	0.0262	0.0305

These data are plotted in Figure 3.31, which uses a logarithm scale since the ratio, not absolute difference, between these two SEs is more likely to be constant. Except for one data point (norethynodrel for lab E), there appears to be a seemingly constant difference between the analytical and simulated-based standard errors.

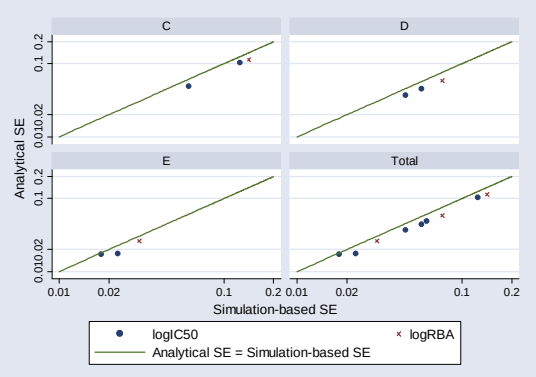

Figure 3.31 Comparison of analytical standard error and simulation-based standard error for logIC_{50} and logRBA

The average of "log(analytical standard error) - log(simulated-based standard error)" is -0.08752 for logIC_{50} and its robust standard error computed by treating labs as clusters is 0.0119, indicating there is statistical evidence (p = 0.02) that the difference is different from zero. This means the use of the analytical standard error would underestimate the standard error by a factor of 0.820 (= $10^{-0.08752}$).

Figure 3.31 includes data points for standard errors of logRBA. The degree of underestimation is similar for logRBA (-0.09118 on log scale, 0.812 on absolute scale).

Using the combined data for $logIC_{50}$ and logRBA, the analytical vs. simulation-based ratio is estimated to be 0.0817.

Using this estimate of the magnitude of underestimation, the 80% prediction intervals for $logIC_{50}$ could be re-computed by multiplying the ratio as a correction factor to each of the input standard errors for the DL method. From the values computed with or without correction for the underestimation given in Table 3.19 and Table 3.20, there was little impact on the end results. The correction is unnecessary.

Table 3.19 Impact of correction for underestimation in SE($logIC_{50}$) on pooled mean and lower and upper limits of 80% prediction intervals

	Estradiol			Norethynodrel		
	Lower limit	Pooled mean	Upper limit	Lower limit	Pooled mean	Upper limit
Without correction	-9.1769	-9.0065	-8.8361	-6.5333	-6.4230	-6.3127
With correction	-9.1756	-9.0050	-8.8344	-6.5287	-6.4156	-6.3025

Table 3.20 Impact of correction for underestimation in SE($logIC_{50}$) on pooled mean and lower and upper limits of 80% prediction intervals

	Norethynodrel		
	Lower limit	Pooled mean	Upper limit
Without correction	-2.6445	-2.6111	-2.5777
With correction	-2.6511	-2.6127	-2.5744

3.6 Setting precision criteria

The method to derive the upper prediction limits for SE($logIC_{50}$) is described in this section since such upper limits made a basis for the upper limits for within-replicate-set variability in % binding. Originally efforts to develop the method for deriving upper limits for SE($logIC_{50}$) and SE(logRBA) were initiated in 2005 based on the perceived importance of precisely estimating $logIC_{50}$ and logRBA. That was the reason for driving the upper limits for SE($logIC_{50}$).

Over time, though, the EDSP has shifted its focus on the qualitative classification of binder vs. non-binder as the main outcome of the receptor binding assay as a screening tool. This meant quantitative measure of potential of a chemical to interact with the receptor in question, i.e., logRBA, is not regarded as important as it was thought to be.

Still, the capacity to accurately and precisely measure logRBA is still considered to be an important feature for a receptor binding assay. Imposing an upper limit for SE(logRBA) for a standard weakly-positive chemical would ensure good precision for logRBA estimates. For this reason the method for deriving the upper limit for SE(logRBA) is described below.

This approach has a limitation in that our estimates of SE(logRBA) is not optimal as discussed in "2.5.2 logRBA" because covariance of $logIC_{50}$ of the standard and $logIC_{50}$ of test chemical, which is a component of var(logRBA), is ignored when SE(logRBA) is estimated. This leaves some uncertainty regarding our description of variation in SE(logRBA) (or to be more precise, variation in log(SE(logRBA)).[103]

One way to keep SE(logRBA) low is to keep both SE($logIC_{50}$(standard)) and SE($logIC_{50}$(positive control)) low. In that sense, imposing upper limits for SE($logIC_{50}$(standard)) and SE($logIC_{50}$(positive control)) still serves our goal.

3.6.1 Standard error of $logIC_{50}$

As discussed in "2.6.1 Standard error of $logIC_{50}$", it is appropriate to describe distribution of SE($logIC_{50}$) on log scale. Logarithm of base 10 was used as a scale. In order to summarize log(SE($logIC_{50}$)) across laboratories, it is necessary to have SE(log(SE($logIC_{50}$))), which is not available from standard output for the DL random effects model used for within-lab, between-run summary of $logIC_{50}$.

In the initial attempt to set up criteria for SE($logIC_{50}$), we used SE(log(SE($logIC_{50}$))) values computed in a simulation study (Feder and Ma, 2005). An alternative is to use the Jack-Knife method of variance estimation. Since performing a simulation study is cumbersome, it would be advisable to use the Jack-Knife method[104] in the future attempts to set up criteria for SE($logIC_{50}$) or similar precision measures such as SE(logRBA). It would be prudent to apply the Jack-Knife method to Task 6 data and compare results with those based on SE(log(SE($logIC_{50}$))) from a simulation.

The initial step for performance criterion setting is selection of acceptable laboratories.[105] For SE($logIC_{50}$), it was determined that labs D and E were acceptable and labs A and C were unacceptable.

SE(log(SE($logIC_{50}$))) was estimated by taking the standard deviation of log(SE($logIC_{50}$)) reported for each simulation round. For estradiol, mean and SE of log(SE($logIC_{50}$)) were as follows.

[103] Dr. Feder proposes to "resolve this with a simulated experiment". If we take this approach, it would be necessary to run a simulated experiment for each chemical. It is not clear whether this needs to be done for each chemical for each run, or whether to be done for each chemical using data from multiple runs and the a common estimate of SE(logRBA) thereby computed would be used for multiple runs. Either way, running a simulated experiment seems to be beyond the usual statistical capacity of a contracting lab. To re-emphasize my point described in footnote 52, by "SE(logRBA)" I am referring to within-run SE(logRBA).
[104] Dr. Feder repeats his points on preference given to delta method over Jack-Knife.
[105] Dr. Feder questions the deletion of a lab (actually two labs here) and asks "How can you estimate lab-to-lab variance based on just two labs?" I agree estimate of variance based on just two labs is unstable, but increasing the number of labs has its own drawbacks of including the labs EDSP has determined to be unacceptable on the basis that are of administrative nature, not so much of statistical nature.

Lab	Mean	SE
D	-1.499	0.243
E	-1.803	0.210

Fitting the DL random effects model to this yields a pooled mean of -1.673 and overall between-lab SE of 0.1587. Intrinsic between-lab SD was 0 (truncated). Using the estimated overall between-lab SE of 0.1587 as an estimated measure of total spread, the upper limit of 95% prediction interval for log(SE(logIC_{50})) from a laboratory drawn from the universe of laboratories that are like labs D and E in terms of SE(logIC_{50}) is computed as

$-1.673 + t_{0.95}$* sqrt(1+1/2)*sqrt(2*0.1587^2)[106]

where $t_{0.95}$ is 6.314 for with degrees of freedom of 1 or 2.920 for degrees of freedom of 2..

Multiplication by 2 (i.e., **2**$*0.1587^2$) is necessary to convert a standard error of pooled mean of estimates from two laboratories to a standard deviation of a future estimate from a single laboratory. [107]

The distributions were computed in a similar manner for logIC_{50} for norethynodrel, and the upper limit was -1.130.

These distributions and upper limits are shown in Figure 3.32, which shows results based on two alternative decisions as to which labs are acceptable. The three panels on the left are based on the decision that labs D and E are acceptable. Those on the right are based on the decision that labs C, D, and E are acceptable. Results for norethynodrel logRBA also are included. As mentioned earlier, the actual decision was to deem labs D and E only as acceptable. Inclusion of lab C as an acceptable lab would have shifted the upper

---

[106] This corrected formula and clarification on $t_{0.95}$ is due to Dr. Feder.

[107] (THE FORMULA AND RESULTING NUMBER IN THIS FOOTNOTE ARE INCORRECT.) In the initial attempt to derive an upper limit for SE(lab-specific logIC_{50}), a standard deviation of log(SE(logIC_{50})) from a laboratory drawn from the universe of laboratories that are like labs D and E was computed using the estimated intrinsic between-lab SD and SE(log(SE(logIC_{50}))) for lab D as a conservative estimate of within-lab variation since SE(log(SE(logIC_{50}))) was greater for lab D than lab E.
Numerically, the computation was

-1.673+ 1.645*sqrt(0^2 +0.243^2) = -1.273

It was realized that it was desirable to use an alternative based on the estimated overall between-lab SE, rather than the combined estimate of overall between-lab variation used in the formula above, as an estimate of overall between-lab variation. The computed standard deviation, -1.273, is slightly more conservative (higher as a number), meaning that its use would lead to higher acceptance of poorly performing labs, than the above-mentioned limit of -1.304. In this particular example, though, the difference was small.

The main reason for preferring the alternative is that the use of greater (greatest) estimate among those of all laboratories combined with the intrinsic between-lab estimate is not necessarily conservative. It also makes the computed upper limit more interpretable statistically. For these reasons, the alternative should be thought of as a default.

limits upwards, but as can be seen in Figure 3.32 the impact of taking the alternative decision would have been relatively small.

In setting these upper limits, the coverage probability of 95% was used. In the examples shown earlier for setting accuracy criteria for $\log IC_{50}$ a lower probability coverage value, *80%,* was used. The choice of the higher coverage probability here was based on an informal judgment regarding the greater uncertainty of the distribution of $\log(SE(\log IC_{50}))$. Data from labs D and E only were used to estimate the distribution. In Figure 3.32 it may be noted that estimates of estradiol $\log IC_{50}$ from lab D would be accepted at a probability of 82%. This probability value would drop to 61% if the coverage probability of 80%, instead of 95%, was used. This probability of 61% seems too low given the fact that D was deemed an acceptable lab and data from it were used in the estimation of the distribution for the acceptable laboratories.[108]

This discrepancy arose because data from only two labs were used for the estimation. If data from a greater number of acceptable laboratories were used for estimating the combined probability distribution, there still would be some labs that would be accepted at a relatively low probability. The proportion of such labs among the labs that were deemed acceptable would be small, and we would not perceive the existence of such labs as alarming. In the present example where only two labs were deemed acceptable, one out of two labs has a seemingly low acceptance probability, and it does seem alarming because we may be rejecting labs like D too often. The derived upper limits may be too stringent and may make it very difficult to qualify many enough labs as acceptable. To avoid such a situation, it seems prudent to use the 95% coverage probability.

The use of this coverage probability still leaves us some capacity to reject labs like lab C, which we judged to be unacceptable. As can be seen in Figure 3.32, estimates from lab C will be rejected at close to 50% for estradiol $SE(\log IC_{50}))$ and at greater than 50% for norethynodrel $SE(\log IC_{50}))$ and norethynodrel SE(logRBA)). As such, the use of the 95% coverage probability seems a reasonable compromise between accepting as many as good labs and rejecting as many as bad labs.

[108] As noted earlier in similar instances, numerical details of the argument in this and following paragraphs are invalid due to the use of incorrect formula. Prediction intervals were narrower than they should be.

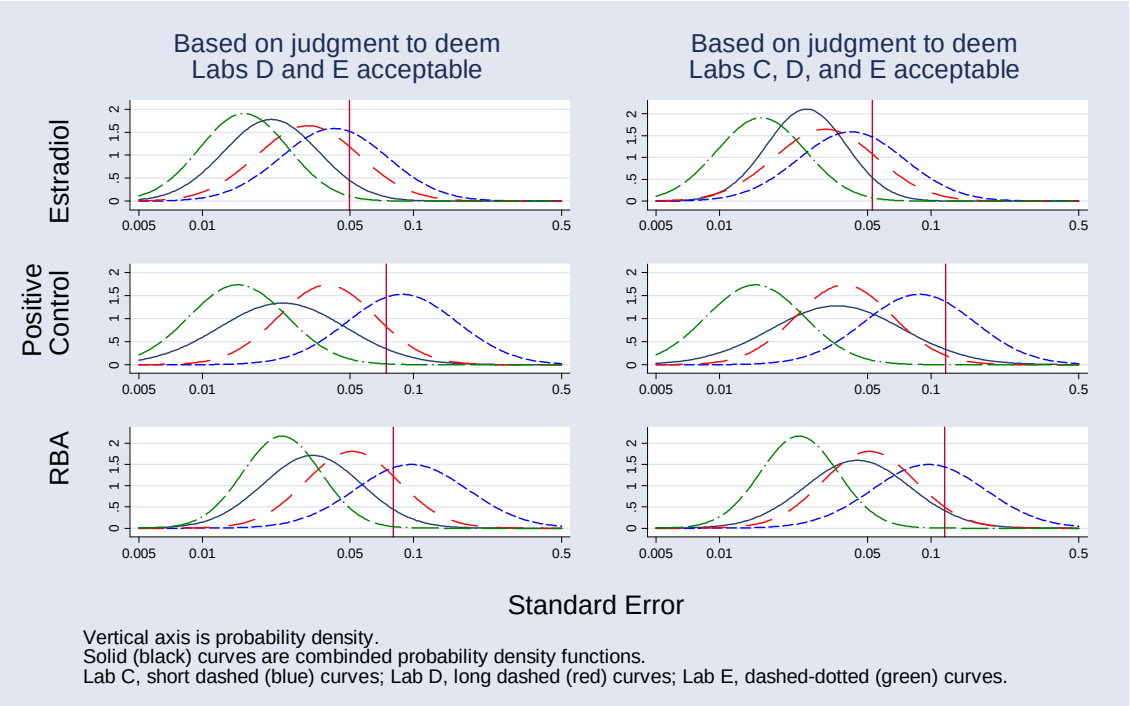

Figure 3.32 Distributions and upper limits based on the 95th percentile of combined distribution for standard error of logIC$_{50}$(estradiol), logIC$_{50}$(norethynodrel), and logRBA(norethynodrel)[109]

The log(SE(logIC$_{50}$)) upper limits for estradiol and norethynodrel were based on simulated log standard errors. Simulated standard errors were observed to be systematically different from observed standard errors for between-lab summary logIC$_{50}$. The observed log(standard error) on average was found to be greater than the simulated log(standard error) by 0.1297, which we can use to correct for the observed vs. simulated difference in log standard errors.

Using the measure of between-lab variation, i.e., the overall between-lab SE, and applying the observed vs. simulated correction, the upper limits for log(SE(logIC$_{50}$))are -1.434 and -1.259 for estradiol and norethynodrel, respectively. By exponentiating these with base of 10, the upper limits for SE(logIC$_{50}$) for estradiol and norethynodrel are computed as 0.036 and 0.074.[110]

Before we move on, let us recapture the discussion on the best method for setting an upper limit for SE(log(SE(logIC$_{50}$))) to be used in the future. First SE(log(SE(lab-specific logIC$_{50}$))) should be estimated using the Jack-Knife method. The distribution of future SE(lab-specific logIC$_{50}$) from an acceptable lab should be described using pooled

[109] The upper limits were incorrect as they were computed using inappropriate formula.
[110] Dr. Feder points what he regards a contradiction between the text and the computational results. In the preceding paragraph, I stated analytical SE was greater than simulated SE, which is taken to be closer to the true SE, by 0.1297. To correct for over-estimation, I subtracted 0.1297 from the observed analytical SE. There seems no contradiction.

between-lab*sqrt(number of labs) as an estimate of overall between-lab SD. The Jack-Knife method has not been applied to Task 6 data yet, but it would be possible to do so.

3.6.2 Standard error of logRBA

The distributions and upper limits were computed for logRBA for norethynodrel in a manner similar to the one described in the previous section. In terms of observed vs. simulated standard errors, the observed log(standard error) on average was found to be *smaller* than the simulated standard error by 0.2725. Applying a correction for this magnitude of underestimation yielded a corrected upper limit of -0.8230. The upper limit for SE(logRBA) for norethynodrel is computed as 0.150 ($=10^{-0.8230}$).

When the alternative measure of between-lab variation, i.e., the overall between-lab SE, is used, the upper limits was computed as -1.011 after observed vs. simulated correction.

3.6.3 Within- and between-replicate-set variability of % binding of radioligand

3.6.3.1 Within -replicate-set variability of % binding of radioligand

The bivariate relationship between within- and between-replicate-set variability of % binding (that is, % binding of radioligand at a given concentration of competitor) and SE(lab-specific $\log IC_{50}$) was described as follows.

$$\log(SD_{\text{between-within-replicate}}(y)) = 1.295 + 0.5894*\log(SE(\text{lab-specific } \log IC_{50}))$$

This equation was obtained by fitting a simple linear regression[111] [112]model to the data consisting of 50 data points, which are depicted as dots in Figure 3.33. Each data point represents a pair of log(SE(lab-specific $\log IC_{50}$)) and $\log(SD_{\text{between-within-replicate}}(y))$. The line corresponding to this estimated equation is shown as a common diagonal line in the two panels in Figure 3.33. Using a common line as opposed to two separate lines for estradiol and norethynodrel, was supported by the lack of evidence for different intercept or slope between the subsets of data belonging to each analyte.

[111] In order to properly account for the correlation among run-specific $SD_{\text{within-replicate}}(Y)$ estimates within the same lab, the use of a "robust" version of simple linear regression is desirable in this situation. When we use a simple linear regression instead, the standard error accompanying the regression coefficient estimates may be incorrect. On the other hand, the use of the simple linear regression allows us to take advantage of a readily accessible procedure to compute predictive interval for the dependent variable for a given value of the predictor variable, which usually is available as an option for a standard linear regression command in statistical software. When actual reported standard errors were compared across these two regression method alternatives, though, the differences appeared to be small. The simple linear regression method was used here because of this and the aforementioned advantage.

[112] Commenting on the approach of using a robust regression explained in the preceding footnote, Dr. Feder states "Why not simply include lab as a block effect in the model. That would introduce equi-correlation within the labs. This is simpler than "robust regression". I take this to mean to include indicators for lab-analyte combinations. This does not seem to work since there is a single value of "x" for each lab-analyte combination as can be seen in Figure 3.33, and there would be no x-y relation after inclusion of the indicators..

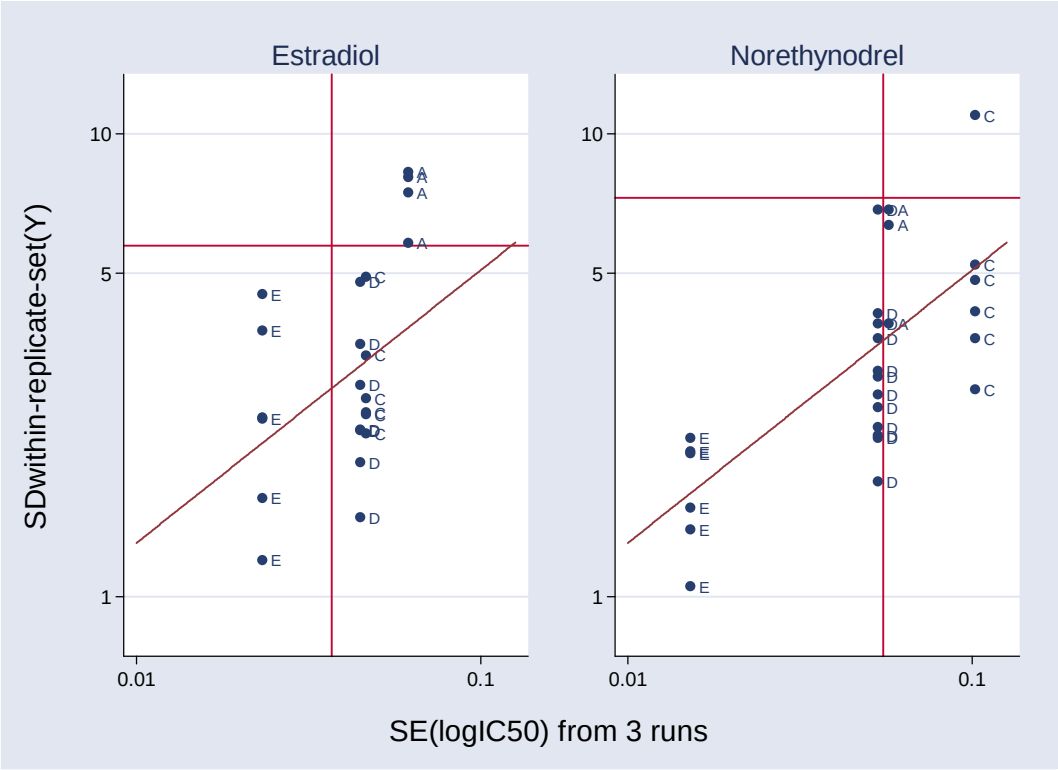

Figure 3.33 Upper bound of the 95% prediction interval for $SD_{within\text{-}replicate\text{-}set}$ variation corresponding to the upper limit set for SE(log(SE(lab-specific $logIC_{50}$)))[113] [114]

How to determine the upper boundary of the one-sided 95% prediction interval for this fitted line is depicted in Figure 3.34. There is a standard procedure to derive a two-sided prediction interval around a fitted line. The shaded area shows the two-sided 90% prediction interval. The upper edge of this interval corresponds to the upper boundary of the one-sided 95% prediction interval, which is quantitatively described as

(upper boundary) =intercept + slope*log(SE(lab-specific logIC50) + STDF *Z0.95 .....3.1[115]

where STDF is the standard error of the forecast supplied as a part of optional output of simple linear regression. STDF is a function of log(SE(lab-specific $logIC_{50}$)), and its computed value is usually available as a part of output from a simple linear regression procedure.

[113] Although we work on $log(SD_{within\text{-}replicate\text{-}set})$, the vertical axis for figures and values in tables in Section 3.6.3 are on the natural scale.
[114] The upper limits in this figure were incorrect as they were computed using inappropriate formula.
[115] Dr. Feder suspects "It looks like the variability of the estimates of intercept and slope are not accounted for." The STDF actually accounts for such a variability. This was illustrated in "Example" on pages 346-7 of "Stata Base Reference Manual Volume 3 Release 8".

The location of the horizontal line in Figure 3.34, which is the upper limit for the dependant variable, is determined by identifying the point at which the upper boundary of the prediction interval and each of two vertical lines for the analyte-specific upper limit for the SE(lab-specific $\log IC_{50}$) cross each other.

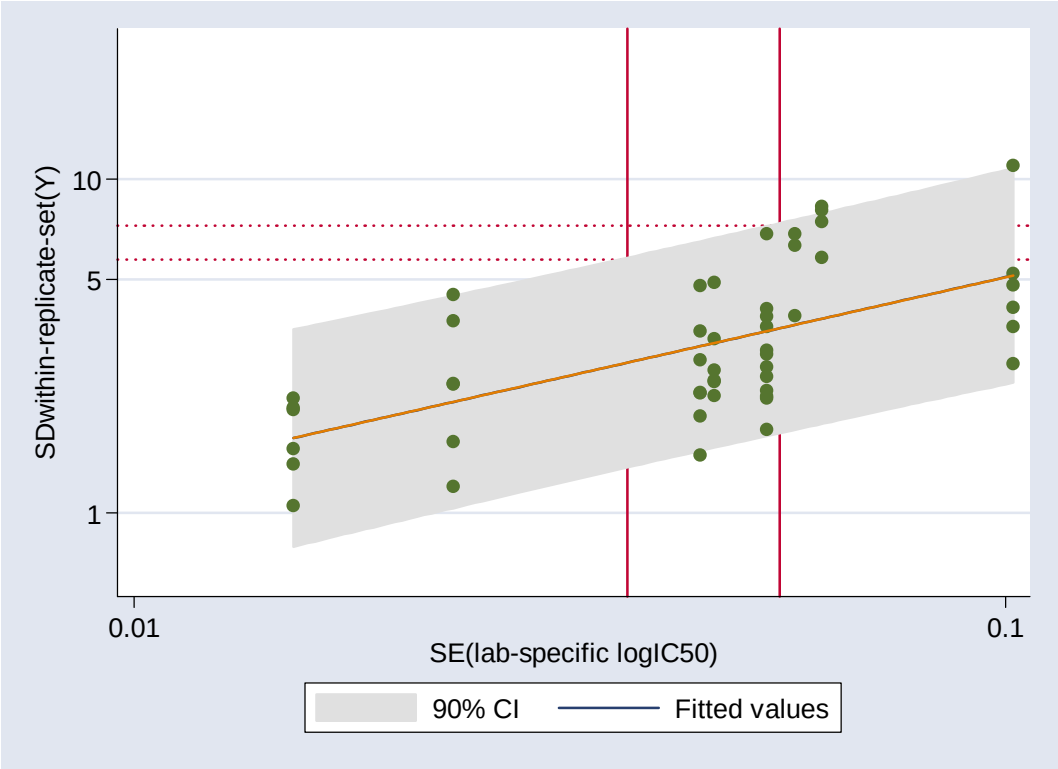

Figure 3.34 How to derive the upper boundary of the 95% prediction interval for $SD_{within\text{-}replicate\text{-}set}$[116]

By changing the vertical width of the shaded area, i.e., percent coverage of prediction interval, we can tweak the acceptance rate of data from a run in terms of whether the within-replicate-set variation in y is low enough. Computed upper limits for several pairs of coverage probability of the prediction interval and the percentile value corresponding to the upper limit of SE(lab-specific $\log IC_{50}$) are shown in Table 3.21

[116] Dr. Feder points out the band shown in this figure "looks funny". He is right. By mistake, instead of correct upper and lower limits that accounts for the contribution of variability due to residual errors and variability of estimates for the intercept and slope, the limits for confidence interval, which account only for the variability of estimates for the intercept and slope, were plotted. The upper limits in this figure also were incorrect as they were computed using inappropriate formula.

Table 3.21 Upper limits of $SD_{within\text{-}replicate\text{-}set}(Y)$ for various levels of prediction interval coverage and acceptance rate for a laboratory like labs D and E[117]

% coverage of prediction interval	Percentile value corresponding to the upper limit of SE(lab-specific $logIC_{50}$)							
	80%		90%		95%		99%	
	Analyte							
	E2*	NE**	E2	NE	E2	NE	E2	NE
80%	~~3.2~~	~~3.7~~	~~3.6~~	~~4.4~~	~~4.1~~	~~5.1~~	~~5.0~~	~~6.8~~
90%	~~3.9~~	~~4.5~~	~~4.4~~	~~5.4~~	~~4.9~~	~~6.2~~	~~6.0~~	~~8.3~~
95%	~~4.5~~	~~5.3~~	~~5.1~~	~~6.3~~	~~5.7~~	~~7.3~~	~~7.1~~	~~9.7~~
99%	~~6.1~~	~~7.1~~	~~6.9~~	~~8.4~~	~~7.7~~	~~9.7~~	~~9.5~~	~~13.0~~

* Estradiol, ** Norethynodrel

By choosing 95% of these, the upper limits for $SD_{within\text{-}replicate\text{-}set}(Y)$ were set to 5.7 and 7.3% for estradiol and norethynodrel, respectively. Choosing these levels is an informal process. We would like to set it high such that we can allow enough labs, but setting it too high results in inclusion of poorly performing labs. It is a balancing act. It is desirable to have many participating laboratories, but as in this case we may have a very small number of labs, only two in this case, that we consider acceptable.

Because of great uncertainty in between-lab variation in SE(lab-specific $logIC_{50}$), it seemed prudent to lean towards leniency. Some levels of percentile values resulted in a cut-off for $SD_{within\text{-}replicate\text{-}set}(Y)$ which seems overly stringent. For instance, when the upper limit for $SD_{within\text{-}replicate\text{-}set}(Y)$ is set at 5.0, a run from lab D would have been almost always rejected even though lab D was one of the two labs deemed acceptable and the data from it were used for deriving the upper limit. When a data point sits at the edge of horizontal line, which is for an upper limit of $SD_{within\text{-}replicate\text{-}set}(Y)$, it means estimates from the lab would be accepted only half the time. Conceptually this indicates that the upper limit was computed based on information from only 1.5 labs. We cannot put much faith in the limit value based on data from so few labs.

This section illustrated how to set an upper limit for within-replicate-set variability estimated using data from a single run. Instead, in some situations it may be desirable to set up an upper limit for within-replicate-set variability estimated using data from three runs.

3.6.3.2 Intrinsic between-replicate-set variability of % binding of radioligand

Intrinsic between-replicate-set variability was measured as lab specific $SD_{between\text{-}replicate\text{-}set}(Y)$ and its relationship with SE(lab-specific $logIC_{50}$) was described. There were only 8 data points (4 laboratories with 2 chemicals tested at each), and the relationship is

[117] The upper limits in this table were incorrect as they were based on incorrect upper limits for SE($logIC_{50}$) computed using inappropriate formula for its prediction interval.

assumed to be common to two chemicals. Figure 3.35 depicts the relationship along with the fitted line and upper limits set for the two variability measures.

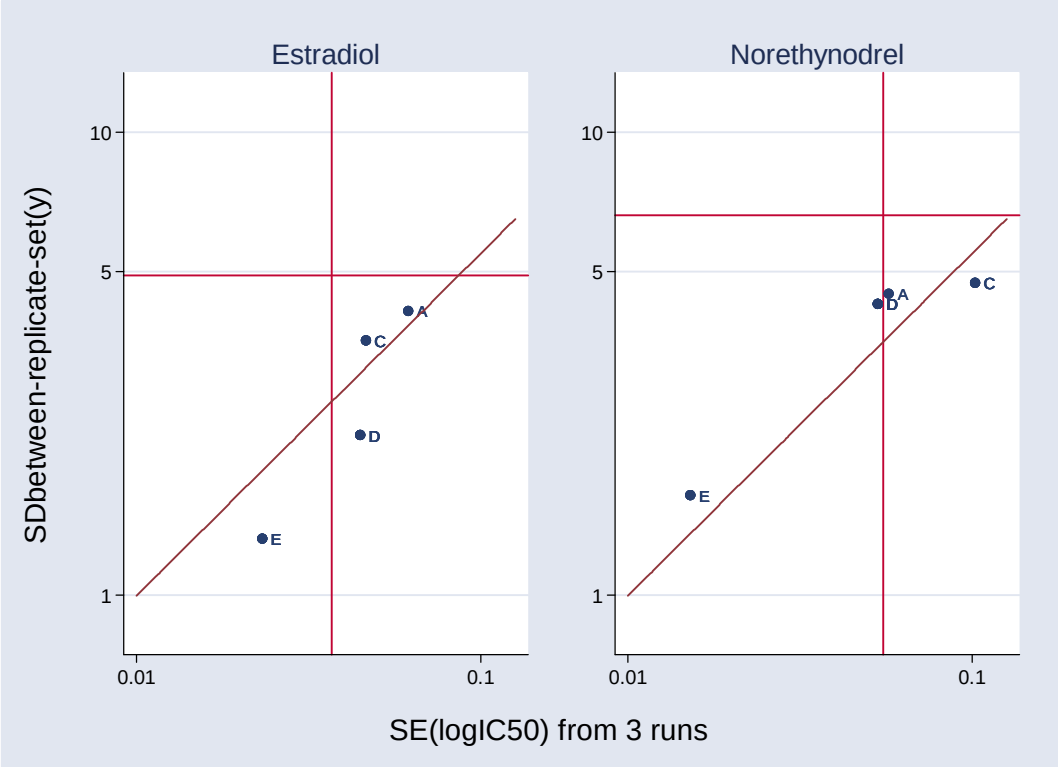

Figure 3.35 Relationship between $SD_{between\text{-}replicate\text{-}set}$ and SE(lab-specific $logIC_{50}$) and the upper limits derived for them[118]

The method for translating the upper limit for SE(lab-specific $logIC_{50}$) to an upper limit for $SD_{between\text{-}replicate\text{-}set}$ is basically the same as the one described for $SD_{within\text{-}replicate\text{-}set}$. The upper limit for SE(lab-specific $logIC_{50}$) was the 95th percentile value computed for the universe of acceptable laboratories. The upper bound of the one-sided 95% prediction interval was used for setting the upper limit for $SD_{between\text{-}replicate\text{-}set}$. How the upper limit changes according to varying percentile values based on which limits are set is shown in Table 3.22. Using 95% both for the acceptance probability of labs in terms of SE($logIC_{50}$) and the coverage probability for prediction interval associated the fitted line, the upper limits for $SD_{between\text{-}replicate\text{-}set}$ were set as 4.1 and 5.5[119] for estradiol and norethynodrel, respectively.

[118] As pointed out by Dr. Feder, the upper limits corresponding to 99% coverage were plotted in this figure. The limits of 4.1% for estradiol and 5.5 for norethynodrel should have been plotted to be consistent with the text. The upper limits in this figure also were incorrect as they were computed using inappropriate formula. Values of $SD_{between\text{-}replicate\text{-}set}$ also were incorrectly computed.
[119] See the preceding footnote.

Table 3.22 Upper limits of $SD_{between\text{-}replicate\text{-}set}$(Y) for various levels of prediction interval coverage and acceptance rate for a laboratory like labs D and E[120]

% coverage of prediction interval	Percentile value corresponding to the upper limit of SE(lab-specific $logIC_{50}$)							
	80%		90%		95%		99%	
	Analyte							
	E2*	NE**	E2	NE	E2	NE	E2	NE
80%	~~2.4~~	~~2.9~~	~~2.9~~	~~3.7~~	~~3.3~~	~~4.4~~	~~4.3~~	~~6.4~~
90%	~~2.8~~	~~3.3~~	~~3.2~~	~~4.1~~	~~3.7~~	~~5.0~~	~~4.8~~	~~7.2~~
95%	~~3.1~~	~~3.7~~	~~3.6~~	~~4.6~~	~~4.1~~	~~5.5~~	~~5.3~~	~~8.0~~
99%	~~3.7~~	~~4.4~~	~~4.3~~	~~5.5~~	~~4.9~~	~~6.6~~	~~6.4~~	~~9.8~~

* Estradiol, ** Norethynodrel

3.6.3.3 Total between-replicate-set variability of % binding of radioligand

Using the $SD_{between\text{-}replicate\text{-}set}$ and $SD_{within\text{-}replicate\text{-}set}$, the variation of the mean of y at a given log(concentration) can be computed as follows.

$$SE_{total\text{-}etween\text{-}replicate\text{-}set} = \sqrt{\frac{(SD_{between\text{-}replicate\text{-}set})^2 + \frac{(SD_{within\text{-}replicate\text{-}set})^2}{\text{Number of replicates}}}{\text{Number of runs}}} \quad \text{.....3.2}$$

For instance, in our default setting of triplicate measurements at each concentration in each run and total of three runs performed, this would be

$$SE_{total\text{-}etween\text{-}replicate\text{-}set} = \sqrt{\frac{(SD_{between\text{-}replicate\text{-}set})^2 + \frac{(SD_{within\text{-}replicate\text{-}set})^2}{3}}{3}} \quad \text{.....3.3}$$

In this setting the mean of y at a given x is a mean of 9 replicates, and that is why in the inside of the square root on the right hand side $(SD_{within\text{-}replicate\text{-}set})^2$ has 9 as a denominator. This mean is a mean of three run-specific means, and that is why $(SD_{between\text{-}replicate\text{-}set})^2$ on the left hand side has 3 as a denominator.

The $SD_{within\text{-}replicate\text{-}set}$ in this expression is a single lab-specific estimate based on data from all available runs rather than a run-specific estimate.

This expression has a practical use. It gives us an idea about how much reduction in $SE_{total\text{-}etween\text{-}replicate\text{-}set}$ we could gain by increasing the number of replicates and/or number

[120] The upper limits in this table were incorrectly computed and so are shown in strikethrough format

of runs. Further, using that estimate of variability reduction we can predict the reduction in SE(lab-specific $\log IC_{50}$) that we can gain by changing the number of replicates or runs in a certain manner.[121]

For each laboratory-analyte combination we can compute $\log(SE_{\text{total-etween-replicate-set}})$ and assess its relationship with log(SE(lab-specific $\log IC_{50}$)). The observed relationship can be used to translate the limits for SE(lab-specific $\log IC_{50}$) into the limits for $SE_{\text{total-etween-replicate-set}}$. The observed relationship between $SE_{\text{total-etween-replicate-set}}$ and SE(lab-specific $\log IC_{50}$) is depicted in Figure 3.36 along with the established upper limits for both dependent and independent variables.

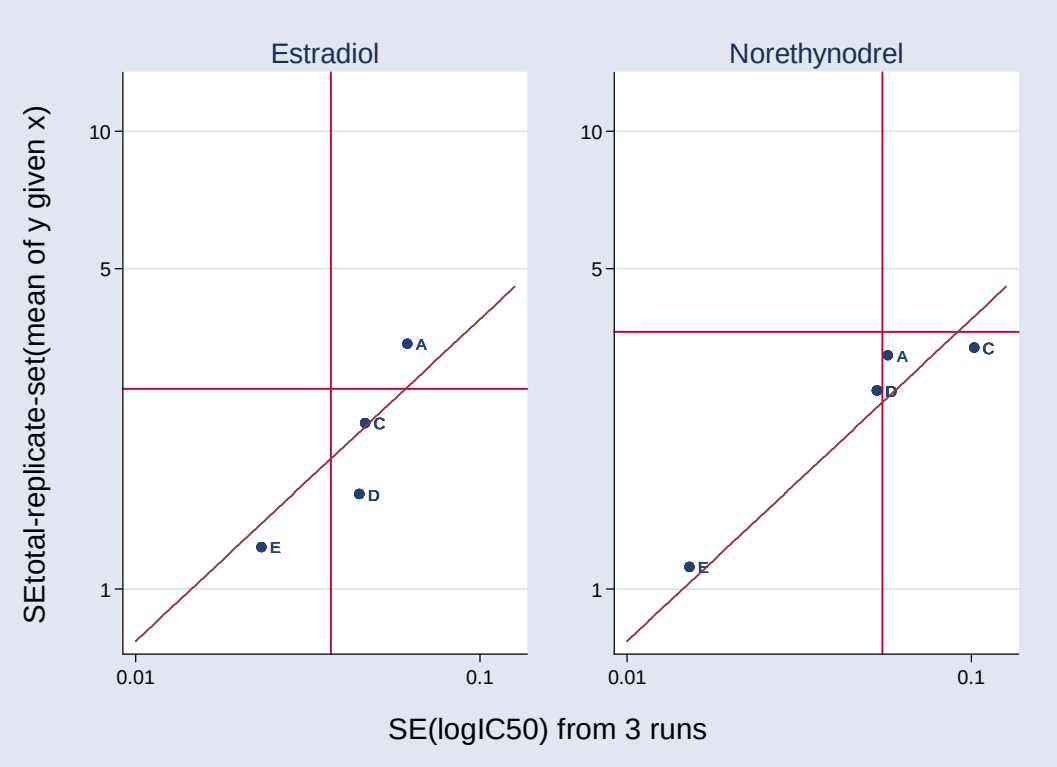

Figure 3.36 Relationship between $SE_{\text{total-between-replicate-set}}$ and SE(lab-specific $\log IC_{50}$) and the upper limits derived for them[122]

A numerical summary of upper limits for varying levels of probability coverage is given in Table 3.23.

[121] A similar expression can be written for between-run and within-run variability for a pooled mean of any of other parameters ($\log IC_{50}$, top plateau, bottom plateau, and Hill slope). Application of this to other parameters may be of more practical advantage. The total between-run variation in the mean y given x is used as an example only because this analysis has been performed in response to the expressed interest of the ESDP in the initial stage of this project.

[122] Values of $SD_{\text{total-between-replicate-set}}$ and upper limits in this figure were incorrectly computed.

A practical use of the expression for $SE_{\text{total-etween-replicate-set}}$ in terms of $SD_{\text{between-replicate-set}}$ and $SD_{\text{within-replicate-set}}$ is illustrated below. The use of this kind of expression was conceptually described in "2.9.2 Uses in assay implementation". Imposing an upper limit for $SE_{\text{total-between-replicate-set}}$ is described as an inequality of the following form.

$$SE_{\text{total-between-replicate-set}} = \sqrt{\frac{(SD_{\text{between-replicate-set}})^2 + \frac{(SD_{\text{within-replicate-set}})^2}{\text{Number of replicate}}}{\text{Number of runs}}} < (\text{Upper limit}) \quad \text{.....3.4}$$

We can now focus on the two components of the SE, i.e., $SD_{\text{between-replicate-set}}$ and $SD_{\text{within-replicate-set}}$.

$$\sqrt{\frac{(SD_{\text{between-replicate-set}})^2 + \frac{(SD_{\text{within-replicate-set}})^2}{\text{Number of replicate}}}{\text{Number of runs}}} < (\text{Upper limit}) \quad \text{.....3.5}$$

The left hand side of this inequality can be manipulated by changing the number of runs and/or number of replicates.[123] Curves for a set of equations corresponding this inequality (change "<" to "=") for various numbers of runs (3, 4, 5) and numbers of replicates (3 and 4) are plotted in Figure 3.37 along with observed values of ($SD_{\text{between-replicate-set}}$, $SD_{\text{within-replicate-set}}$). If a data point is below a certain curve, the inequality corresponding to the curve holds or the point is acceptable in terms of the imposed upper limit included in the inequality. For instance, data points corresponding to labs C, D, and E are below the thick black curve in the estradiol panel, indicating they met the criteria using 3 runs in triplicate. On the other hand, the data point for lab A was above the curve, meaning it was not accepted when using 3 runs in triplicate are used. Lab A using 3 runs in triplicate is disqualified for estradiol because the data point representing it is above the black thick curve. If lab A used 5 runs in triplicate or quadruplicate, it would be able to generate low enough SE($\log IC_{50}$). The black thick curves correspond to the horizontal lines in Figure 3.36.

In general, we would not favor giving special treatment to a particular lab, and so this example is not realistic. In some other circumstances, e.g., where an agency is finding it extremely difficult to find any laboratories that meet certain pre-specified precision requirement, the method described above could be useful. If this type of difficulty is faced, it may be justifiable to tweak the experimental protocol and increase the default number of runs, for instance, to allow enough laboratories to produce data that meet the pre-specified precision requirement.

[123] Dr. Feder repeats his criticism on an approach of this kind. My response may be found in footnote 19.

Table 3.23 Upper limits of $SD_{total\text{-}between\text{-}replicate\text{-}set}$(Y) for various levels of prediction interval coverage and acceptance rate for a laboratory like labs D and E[124]

% coverage of predictive interval	Percentile value corresponding to the upper limit of SE(lab-specific $logIC_{50}$)							
	80%		90%		95%		99%	
	Analyte							
	E2*	NE**	E2	NE	E2	NE	E2	NE
80%	~~1.7~~	~~2.1~~	~~2.0~~	~~2.6~~	~~2.3~~	~~3.1~~	~~3.0~~	~~4.3~~
90%	~~1.9~~	~~2.3~~	~~2.2~~	~~2.8~~	~~2.5~~	~~3.4~~	~~3.2~~	~~4.8~~
95%	~~2.1~~	~~2.5~~	~~2.4~~	~~3.0~~	~~2.7~~	~~3.6~~	~~3.5~~	~~5.2~~
99%	~~2.4~~	~~2.9~~	~~2.8~~	~~3.5~~	~~3.2~~	~~4.2~~	~~4.1~~	~~6.1~~

* Estradiol, ** Norethynodrel

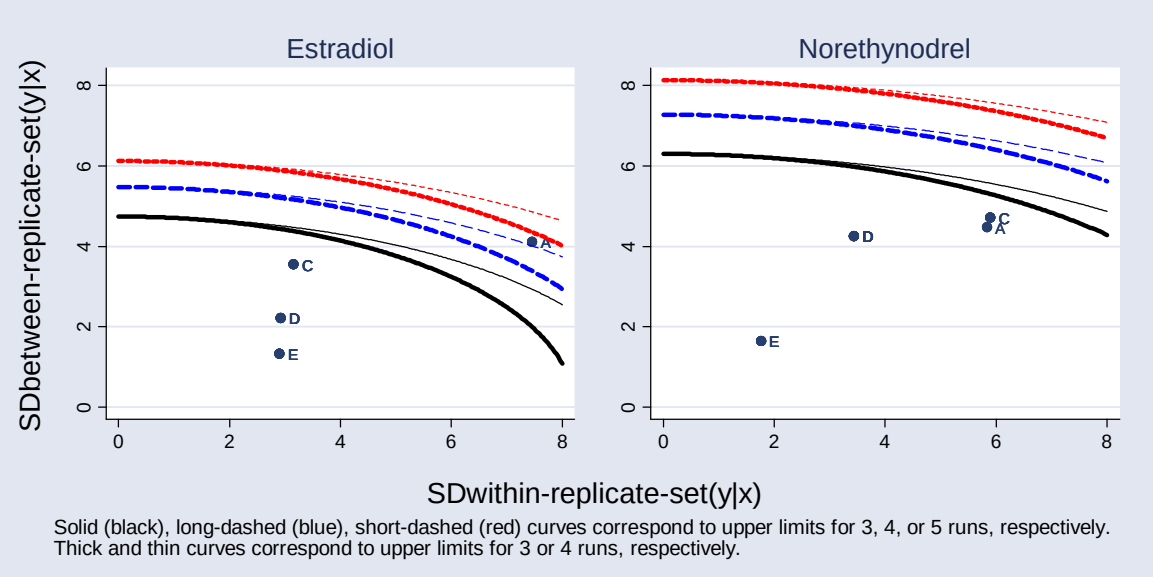

Figure 3.37 Relationship between $SE_{total\text{-}between\text{-}replicate\text{-}set}$ and SE(lab-specific $logIC_{50}$) and the upper limits derived for them[125]

3.6.3.4 Alternative method for derivation of upper limits

The development of the methods described in Section 3.6.3 was driven by the interest of the EDSP expressed in 2005 to keep SE($logIC_{50}$) at a low level. To achieve that end, the $SD_{within\text{-}replicate\text{-}set}$ was to be quantified after each run and data from the runs with low enough $SD_{within\text{-}replicate\text{-}set}$ were to be deemed acceptable.

The EDSP's position on the importance of $logIC_{50}$ has been changed since, and it no longer places a high priority on the need to keep SE($logIC_{50}$) low.[126] [127]Given that change,

[124] The upper limits in this table were incorrectly computed and so are shown in strikethrough format
[125] Values of SDtotal-between-replicate-set and upper limits in this figure were incorrectly computed.
[126] Keeping SE($logIC_{50}$) low for the standard and positive control is expected to result in low SE(logRBA) for the positive control. This itself still may form a basis for the continued use of the limits for within-replicate-set variability based on the limit for SE($logIC_{50}$). In computing SE(logRBA), we need to ignore the covariance of $logIC_{50}$ for the standard and that for the positive control because of lack of easy way to

the method for deriving the upper limit for within-replicate-set variability and related variability measures may be completely overhauled.

To reflect the decreased emphasis on $logIC_{50}$, we may disconnect $SD_{within\text{-}replicate\text{-}set}$ from $SE(logIC_{50})$ and regard it as a variability measure of its own merit. Conceptually, a reduction in $SD_{within\text{-}replicate\text{-}set}$ would result in reduction in the standard deviation of any of the four Hill equation parameters, i.e., not only $logIC_{50}$ but also the top, bottom, and slope parameters. Based on this general idea, an alternative method for derivation of upper limits for $SD_{within\text{-}replicate\text{-}set}$ is conceived and described below.

We will treat $log(SD_{within\text{-}replicate\text{-}set})$ as a parameter. Observed distributions of $log(SD_{within\text{-}replicate\text{-}set})$ for existing acceptable labs are described, and then we estimate the distribution of $log(SD_{within\text{-}replicate\text{-}set})$ for the universe of acceptable labs. Based on the distribution for the universe of acceptable labs, a level below which estimates from most of acceptable labs would fall may be determined.

Specifically, $log(SD_{within\text{-}replicate\text{-}set})$ will be computed after each run using the one-way random effects ANOVA. Following a procedure similar to what we applied to a Hill equation parameter, e.g., top plateau level parameter, we can summarize $log(SD_{within\text{-}replicate\text{-}set})$ across runs within a lab, then across labs using the DL random effects model. Once between-lab and within-lab variation is quantified for $log(SD_{within\text{-}replicate\text{-}set})$ it is straightforward to establish a one-sided prediction interval for $log(SD_{within\text{-}replicate\text{-}set})$ using the procedures applied to other Hill equation parameters described in "2.5 Deriving accuracy criteria" and exemplified in Sections 3.2-3.5. The only main difference would be that limits for $log(SD_{within\text{-}replicate\text{-}set})$ will be based on one-sided prediction intervals while the derivation of the accuracy criteria was based on two-sided prediction interval.

In order to apply the DL method to $log(SD_{within\text{-}replicate\text{-}set})$, we will need SE $log(SD_{within\text{-}replicate\text{-}set})$. Regular output or interim results of the one-way random effects ANOVA does not include any numerical information that can be used directly to estimate $SE(log(SD_{within\text{-}replicate\text{-}set}))$. The Jack-Knife method may be used to estimate $SE(log(SD_{within\text{-}replicate\text{-}set}))$ for each run, using a replicate set as a unit of observation to compute pseudovalues.[128]

The use of this alternative method would treat the $log(SD_{within\text{-}replicate\text{-}set})$ as a parameter in a manner comparable to that used for other Hill equation parameters and would provide a unified look for the entirety of this performance criterion derivation exercise. As done for

---

estimate it. As a result, there is an increased level of uncertainty for the estimated distribution of log(SE(logRBA)), and it may not be prudent to fully trust the upper limits derived using that distribution. The distribution of $log(SE(logIC_{50}))$ is more reliably estimated, and upper limits for $log(SE(logIC_{50}))$ thus may be more trustable.

[127] In commenting on the preceding footnote, Dr. Feder notes "We can estimate correlation within run" and offers a sketch of how the covariance could be estimated on page 143 of his hand-written comments (Feder, 2007b).

[128] Dr. Feder repeats his recommendation that delta method be used in place of Jack-Knife in this context.

the upper limits for SE($\log IC_{50}$) and SE(logRBA), the derived limits would be back-transformed to the natural scale and expressed as limits for $SD_{\text{within-replicate-set}}$.[129]

Similar procedures may be applied to $\log(SD_{\text{between-replicate-set}})$ and $\log(SD_{\text{overall-between-replicate-set}})$ in order to derive upper limits for $SD_{\text{between-replicate-set}}$ and $SD_{\text{between-replicate-set}}$, respectively.

[129] It may be feasible to use a kind of hierarchical random effects ANOVA model and based on it directly construct an upper limit for $\log(SD_{\text{within-replicate-set}})$, avoiding the use of the DL random effects model. It is beyond the scope of this report to discuss such an approach. The DL method-based approach described here has an appeal of relatively easy implementation and comparability with the procedures used for other parameters.

4 Appendix

4.1 Alternative definition of a unit in partitioning variation in % binding

A decision to use a replicate set as unit for partitioning variation in y given x was explained in "2.3.1 Definition of ". There is an alternative to this choice a unit. Justification for our choice is given below.[130]

An alternative way to define the unit of interest is to define a unit as each run as in the following table. There are 3 units, each having 21 (triplicates times 7 levels of x) observations.

Table 4.1 Preparation of receptor binding data for computation of within-unit and between-unit variance: unit specification by run alone

	Run								
	1			1			1		
	Replicate			Replicate			Replicate		
x	1	1	1	1	1	1	1	1	1
x_1	y- $\overline{y}_1$	y- $\overline{y}_1$	y- $\overline{y}_1$	y- $\overline{y}_1$	y- $\overline{y}_1$	y- $\overline{y}_1$	y- $\overline{y}_1$	y- $\overline{y}_1$	y- $\overline{y}_1$
x_2	y- $\overline{y}_2$	y- $\overline{y}_2$	y- $\overline{y}_2$	y- $\overline{y}_2$	y- $\overline{y}_2$	y- $\overline{y}_2$	y- $\overline{y}_2$	y- $\overline{y}_2$	y- $\overline{y}_2$
x_3	y- $\overline{y}_3$	y- $\overline{y}_3$	y- $\overline{y}_3$	y- $\overline{y}_3$	y- $\overline{y}_3$	y- $\overline{y}_3$	y- $\overline{y}_3$	y- $\overline{y}_3$	y- $\overline{y}_3$
x_4	y- $\overline{y}_4$	y- $\overline{y}_4$	y- $\overline{y}_4$	y- $\overline{y}_4$	y- $\overline{y}_4$	y- $\overline{y}_4$	y- $\overline{y}_4$	y- $\overline{y}_4$	y- $\overline{y}_4$
x_5	y- $\overline{y}_5$	y- $\overline{y}_5$	y- $\overline{y}_5$	y- $\overline{y}_5$	y- $\overline{y}_5$	y- $\overline{y}_5$	y- $\overline{y}_5$	y- $\overline{y}_5$	y- $\overline{y}_5$
x_6	y- $\overline{y}_6$	y- $\overline{y}_6$	y- $\overline{y}_6$	y- $\overline{y}_6$	y- $\overline{y}_6$	y- $\overline{y}_6$	y- $\overline{y}_6$	y- $\overline{y}_6$	y- $\overline{y}_6$
x_7	y- $\overline{y}_7$	y- $\overline{y}_7$	y- $\overline{y}_7$	y- $\overline{y}_7$	y- $\overline{y}_7$	y- $\overline{y}_7$	y- $\overline{y}_7$	y- $\overline{y}_7$	y- $\overline{y}_7$

A statistically natural name for the within-variation calculated using this definition of unit may arguably be $SD_{within\text{-}run}$. [131] This measure includes not only the variation within a

[130] Dr. Feder request additional explanation as to what this section is intended for. The reason is, as stated, to give justification for how I partitioned total variability of individual y measurements. The procedure of my choice is partitioning the variability to "within-replicate" and "between-replicate". The within-replicate variability corresponds to, as I learned from interaction with participants of conference calls on inter-lab study for ER binding assay, what experimentalists refer informally as "within-run variation". Biostatisticians, though, would reserve the term "within-run" variation for the "within" components of the alternative procedure for variability partitioning. As such, the purpose of this section is to clearly show that which kind of partitioning I was using and to justify why it was chosen. I stated the main reason for this choice was a good analogy to a standard between-subject set-up. I should have mentioned another, arguably more important, reason is that between-replicate-set variation could be computed from data from a single run. With data from a single run, the "wiggle" component could not be evaluated and so the alternative procedure does not work if we need an estimate of "within" variability from a single run. As explained in Aoki (2007b), such an estimate that can be obtained with data from a single run was initially desired by EDSP.

triplicate set, but also how y^* varies across x levels within a run. The $SD_{between\text{-}replicate\text{-}set}$ derived using this definition, includes a parallel shift of the entire curve from the curve constructed by joining the means at each x.

The two alternative pairs of definitions for within- and between-replicate-set variabilities arise from the fact that y^* actually has three sources of variation depicted in Figure 4.1.

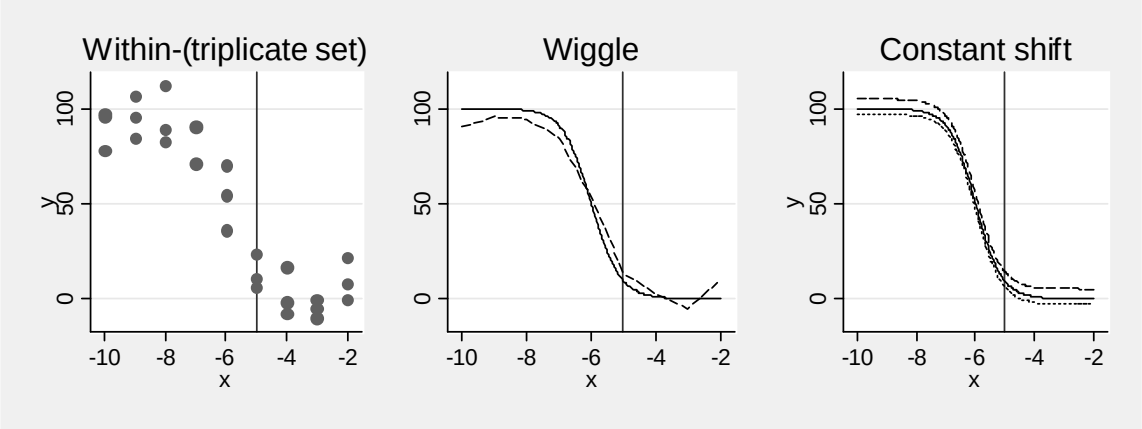

Figure 4.1 Three sources of overall between-run variation

In this figure, how y^* varies across x levels within a run is tentatively labeled as "wiggle" and is illustrated in the middle panel. Depending upon which definition of "unit" is used (i.e., replicate set or run), this "wiggle" variation could be partitioned into either within-unit or between-unit variation. If the unit is the replicate set, wiggle becomes part of between-unit variation; if the unit is the run, wiggle becomes part of within-unit variation.

It has been decided to combine the "wiggle" variation and "constant shift" variation, separating the former from within-triplicate-set variation. The main reason to do this is that there is natural analogy between the within-triplicate-set variation in our receptor binding assay setting and the within-unit variation in the standard between-subject setup in that both deal only with measurement error at the lowest level of organization.

4.2 Improved correction for the number of runs performed

While finalizing this report, an important issue related to the method described in "2.2.2 Correction for the number of runs performed" was noted. The method deals with varying numbers of runs across labs (e.g., 3 runs worth of data are available from on lab while 6 runs worth of data are available from the other) probably was not optimal.

The proposed and actually used method was to employ the following formula to generate a hypothetical standard error value that would correspond to data from three runs.

[131] What we defined as $SD_{within\text{-}replicate\text{-}set}$ tended to be called $SD_{within\text{-}run}$ by experimentalists.

$$SE(\hat{\theta}_{\text{R for lab i}}) = \sqrt{\frac{1}{\left(\sum_{j=1}^{k_i} \frac{1}{SE(\hat{\theta}_{ij})^2 + \hat{\tau}_i^{\,2}}\right)\frac{3}{k_i}}} \quad \text{.....2.20}$$

Standard error values corrected in this manner are used when combining lab-specific pooled means across labs. This forces the laboratories to be comparable in terms of the numbers of runs.

In hindsight, this solved one problem of imbalance in the number of runs, but created another. The problem is that the lab-specific means were still computed on the original full data, not from just three runs, and they were on average less variable than the lab-specific means computed on the data from three runs only. In the data we used as an example, the numbers of runs were 3 or more. It would have been desirable to use the mean number of runs per lab instead of a fixed number of runs (i.e., 3) in the following manner.

$$SE(\hat{\theta}_{\text{R for lab i}}) = \sqrt{\frac{1}{\left(\sum_{j=1}^{k_i} \frac{1}{SE(\hat{\theta}_{ij})^2 + \hat{\tau}_i^{\,2}}\right)\frac{\sum_{i=1}^{N} k_i / N}{k_i}}} \quad \text{.....4.1}$$

For instance, for norethynodrel there were data from 3, 6, 12, and 6 runs for labs A, C, D, and E, respectively. In stead of $3/k_i$ on the right hand-side of this equation as a correction factor, we would use $6.25/ k_i$ (i.e., $\sum_{i=1}^{N} k_i / N = (3+6+12+6)/4 = 6.25$).

The use of Equation 4.1 in place of Equation 2.20 is recommended in the future. It would be advisable that the analyses presented in this report be repeated using Equation 4.1 instead of Equation 2.20.[132]

The pooled estimate of within-lab SD previously described,

$$SD_{\text{within-lab}}(\hat{\theta}_{\text{i}}) = \sqrt{\left(\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}}\right) - \hat{\tau}^2} \quad \text{.....2.23}$$

where k is the total number of laboratories.

[132] Please see Section "2.3.1.2. Correction for the number of runs pre lab" in Aoki (2007b) for additional discussion on how is to be computed from data with varying number of runs (≠ 3) per lab.

as computed following the between-lab summary using Equation 4.1 is for $\hat{\theta}_{\text{R for lab i}}$ based on $k_m = \sum_{i=1}^{N} k_i / N$ runs (k_m is the average number of runs per lab). In order to compute the $SD_{\text{within-lab}}$ based on a different number of runs, 3 (out default) for example, is

$$SD_{\text{within-lab}}(\hat{\theta}_{\text{i}} \text{ based on 3 runs}) = \sqrt{\left(\left(\frac{k}{\sum_{i=1}^{k}\frac{1}{SE(\hat{\theta}_i)^2+\hat{\tau}^2}}\right)-\hat{\tau}^2\right)\frac{k_m}{3}} \quad \text{.....4.2}$$

$SD_{\text{within-lab}}$ for a single run is

$$SD_{\text{within-lab}}(\hat{\theta}_{\text{i}} \text{ based on a single run}) = \sqrt{\left(\left(\frac{k}{\sum_{i=1}^{k}\frac{1}{SE(\hat{\theta}_i)^2+\hat{\tau}^2}}\right)-\hat{\tau}^2\right)k_m} \quad \text{.....4.3}$$

Dr. Feder recommends alternative formulae for Equations 4.2 and 4.3.

$$SD_{\text{within-lab}}(\hat{\theta}_{\text{i}} \text{ based on 3 runs}) = \sqrt{\left(\left(\frac{k}{\sum_{i=1}^{k}\frac{1}{SE(\hat{\theta}_i)^2+\hat{\tau}^2}}\right)-\hat{\tau}^2\right)\frac{k_m}{3}} \quad \text{..... 4.2Alternative to 4.2}$$

$$SD_{\text{within-lab}}(\hat{\theta}_{\text{i}} \text{ based on a single run}) = \sqrt{\left(\frac{k}{\sum_{i=1}^{k}\frac{1}{SE(\hat{\theta}_i)^2 k_i+\hat{\tau}^2}}\right)-\hat{\tau}^2} \quad \text{..... Alternative to 4.3}$$

These formulae imply the following two alternative formulae for the $SD_{\text{total-between-lab}}$.

$$SD_{\text{within-lab}}(\hat{\theta}_{\text{i}} \text{ based on 3 runs}) = \sqrt{\left(\left(\frac{k}{\sum_{i=1}^{k}\frac{1}{SE(\hat{\theta}_i)^2+\hat{\tau}^2}}\right)-\hat{\tau}^2\right)\frac{k_m}{3}} \quad \text{.....4.2} \qquad 4.4$$

$$SD_{\text{total-between-lab}}(\hat{\theta}_i \text{ based on a single run}) = \sqrt{\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 k_i + \hat{\tau}^2}}} \quad \text{.....4.5}$$

These make better sense than the ones I was using. In place of the last two formulae, I was using the following two although that was not properly documented in the draft report.

$$SD_{\text{total-between-lab}}(\hat{\theta}_i \text{ based on 3 runs}) = \sqrt{\hat{\tau}^2 + \left(\left(\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}} \right) - \hat{\tau}^2 \right) \frac{k_m}{3}} \quad \text{.....4.6}$$

$$SD_{\text{total-between-lab}}(\hat{\theta}_i \text{ based on a single run}) = \sqrt{\hat{\tau}^2 + \left(\left(\frac{k}{\sum_{i=1}^{k} \frac{1}{SE(\hat{\theta}_i)^2 + \hat{\tau}^2}} \right) - \hat{\tau}^2 \right) k_m} \quad \text{.....4.7}$$

That is, pooled within-lab variance was calculated and it was added to intrinsic between-lab variance in order to reconstruct total between-lab variance. The use of equations 4.4 and 4.5 would eliminate this redundant step.

A preliminary investigation into the effects of the suboptimal correction that was employed mostly in this report was performed using equations 4.2, 4.3, 4.6, and 4.7. Its results are shown in Table 4.2 below.

For all parameter-analyte combinations, the choice of a correction method made little difference in the estimates of overall mean. SE(overall mean) values also were not affected substantially except for the bottom plateau parameter. SE(overall mean), at any rate, is not directly used in derivation of performance criteria.[133]

Some of the within- and between-variability measures, though, are substantially affected. Changes in intrinsic between-lab SD and within-lab SD directly affect performance criteria to be derived based on them. Overall between-lab SD is a measure that combined the between-lab SD and within-lab SD for the lab-specific summary estimates based on three runs. The within-lab SD values changed little no matter which of the original or improved correction factor was used. The use of the improved correction factor increased the intrinsic between-lab SD estimates (and thereby increased overall between-lab SD

[133] Dr. Feder comments as follows. "I disagree. SE(overall mean) is used in setting the performance criteria for the accuracy formulae, e.g., Eq. 2.17, 2.18, 2.40." Actually I was not using Eq. 2.17 and 2.18 directly. Equation 2.40 includes $SD(\hat{\theta})$, which was computed using a formula in the form of Equation 4.6 or 4.7 (as mentioned in the paragraph preceding Eq. 4.6, this was not properly documented in the draft report). If we use a formula for total variability in the form of Equation 4.4 or 4.5 as Dr. Feder would have done, SE and SD are more closely connected and SE(overall mean) is almost directly used in setting performance criteria.

estimates, too) for the top plateau and Hill slope parameter. This indicates that the limits derived for these parameters using the original correction factor may have been overly stringent. It may account at least partially for the previously-noted observation that the prediction interval for norethynodrel's Hill slope seemed to narrow (Figure 3.26 and the text immediately after the figure). Since the EDSP is using "three out of five" rule, which seems to be fairly lenient (see discussion on this in Section "3.5.5 logRBA"), it seems unlikely that somewhat stringent limits have made the coverage probability of the limits too low.

Overall, the use of the suboptimal correction factor did not seem to have profoundly compromised the utility of the derived performance criteria. Nonetheless, the use of the improved correction factor is recommended in the future since the potential adverse consequences of using the suboptimal correction factor could be inferred theoretically as well as empirically from the result of preliminary investigation.

Table 4.2 Comparisons of across-lab summary results based on the original and improved correction factors

Param-eter	Analyte	"Number of run" correction method	Overall mean $\hat{\theta}_{\bar{R}}$	SE(overall mean) $SE(\hat{\theta}_{\bar{R}})$	Overall between-lab SD $SD_{lab}(\hat{\theta})$	Intrinsic between-lab SD $\hat{\tau}_{lab}$	Within-lab SD $SD_{\text{within-lab}}(\hat{\theta}_i)$	Intraclass correlation	Between/ within ratio	Hetero-geneity *p*-value
$logIC_{50}$	Estradiol	Original	-8.997	0.0600	0.1200	0.1112	0.0450	0.859	2.47	$1*10^{-05}$
		Improved	-8.998	0.0596	0.1237	0.1150	0.0456	0.864	2.52	$2*10^{-11}$
	Norethynodrel	Original	-6.454	0.0964	0.1928	0.1826	0.0620	0.897	2.95	$2*10^{-10}$
		Improved	-6.457	0.0954	0.1962	0.1857	0.0633	0.896	2.93	$2*10^{-21}$
Top plateau	Estradiol	Original	100.23	1.28	**2.55**	**1.46**	2.09	0.326	0.70	0.21
		Improved	100.16	1.32	**3.13**	**2.09**	2.33	0.446	0.90	0.02
	Norethynodrel	Original	99.01	0.92	**1.84**	**0.00**	1.84	0.000	0.00	0.44
		Improved	98.75	0.93	**2.34**	**1.25**	1.97	0.287	0.63	0.13
Bottom Plateau	Estradiol	Original	-1.21	**0.37**	0.73	0.00	0.73	0.000	0.00	0.82
		Improved	-1.21	**0.25**	0.73	0.00	0.73	0.000	0.00	0.59
	Norethynodrel	Original	-2.06	**0.52**	1.03	**0.00**	1.03	0.000	0.00	0.67
		Improved	-2.04	**0.38**	1.08	**0.22**	1.05	0.042	0.21	0.36
Hill Slope	Estradiol	Original	-0.97	0.020	**0.040**	**0.007**	0.040	0.029	0.17	0.38
		Improved	-0.98	0.022	**0.055**	**0.031**	0.046	0.315	0.68	0.09
	Norethynodrel	Original	-0.95	0.032	**0.064**	**0.035**	0.053	0.310	0.67	0.24
		Improved	-0.96	0.033	**0.079**	**0.053**	0.058	0.453	0.91	0.03
logRBA	Norethynodrel	Original	-2.57	0.0340	0.0680	0.0527	0.0430	0.601	1.23	0.05
		Improved	-2.57	0.0340	0.0696	0.0544	0.0433	0.613	1.26	0.03

Values in bold indicate instances where the two versions of the correction factor made noticeable difference.

4.3 Statistical software

Stata version 8 was used for the analyses described in Sections “2 Statistical Methods" and “3 Examples”. Statistical methods used include the DL random effects model (implemented via a user-defined module called “meta”), random effects one-way ANOVA (“loneway” command), and OLS with or without robust SE estimation (“regress” command with a cluster option).

References

1) Aoki Y. (2007a) Report On STATISTICAL METHODS FOR EVALUATING VARIABILITY IN AND SETTING UP PERFORMANCE CRITERIA FOR RECEPTOR BINDING ASSAYS: DRAFT (Task1_report_20070413.doc), April 6, 2007
2) Aoki Y. (2007b) Response to comments by Dr. Paul Feder On Draft Report On STATISTICAL METHODS FOR EVALUATING VARIABILITY IN AND SETTING UP PERFORMANCE CRITERIA FOR RECEPTOR BINDING ASSAYS
3) Aoki Y. (2007c) Guidance on revision of performance criteria numbers (perf_criteria_rev_20070608.doc), June 8, 2007
4) DerSimonian R, Laird N. (1986) Meta-analysis in clinical trials. Control Clin Trials. 1986 Sep;7(3):177-88.
5) Feder PI. (2007a) Comments prepared as an electronic file titled "CommentsonYutakaAokiReport_05212007.doc", May 22, 2007
6) Feder PI. (2007b) Hand-written comments recorded in "YutakaAokiannotatedreport_05312007.pdf", May 31, 2007
7) Feder PI. (2007c) Tolerance Intervals: Alternative Performance Criteria by which to Determine Acceptability of Competitive Binding Assays Results, June 4, 2007
8) Feder PI, Ma ZJ. (2005) Draft Report on SENSITIVITY OF ESTIMATES OF IC_{50} AND RELATIVE BINDING AFFINITY, AND THEIR STANDARD ERRORS, TO STATISTICAL ANALYTICAL METHOD. Battelle. Columbus, OH
9) Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ. 2001 Jun 16;322(7300):1479-80.
10) Motulsky H, Christopoulos A. (2003) Fitting Models to Biological Data using Linear and Nonlinear Regression. GraphPad Software, Inc. San Diego, CA.
11) National Research Council. Combining information: statistical issues and opportunities for research. Washington, DC: National Academy Press, 1992.
12) Stata Corp. (2003) Stata Base Reference Manual Volume 3 Release 8.
13) Takkouche B, Cadarso-Suarez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. Am J Epidemiol. 1999 Jul 15;150(2):206-15.