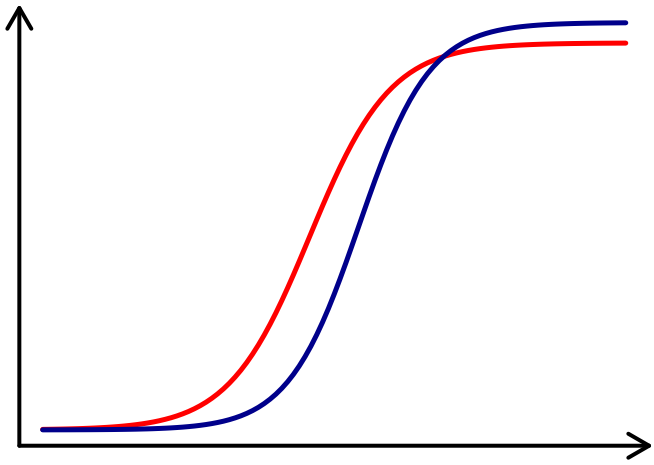
UNIVERSITÉ DE LIÈGE

Faculté de Médecine

Contribution to Statistical Similarity Testing in Potency Assays



Thèse présentée en vue de l'obtention du grade de
Docteur en Sciences Biomédicales et Pharmaceutiques, par

Perceval Sondag

Sous la direction de

Prof. Anne-Françoise Donneau & Dr. Pierre Lebrun

2020

UNIVERSITÉ DE LIÈGE

Faculty of Medicine

Department of Public Health Sciences

Contribution to Statistical Similarity Testing in Potency Assays

Perceval Sondag

Jury members:

Dr. Julien Hanson	President
Dr. Eric Ziemons	Secretary
Prof. Anne-Françoise Donneau	Supervisor
Dr. Pierre Lebrun	Co-supervisor
Dr. Bruno Boulanger	
Prof. Helena Geys	
Prof. Fulvio De Santis	

2020

This page is intentionally left blank

Abstract

Potency assays measure the biological activity or binding affinity of a substance. These measurements are made throughout the entire pharmaceutical product development and supply process, with different objectives at each phase. A common measure of potency is the C_{50}, which reflects the concentration that elicits a response equal to 50% of the response range. The C_{50} is estimated by modeling the substance response as a function of concentration, often using a four-parameter logistic (4PL) function.

The C_{50} estimation may be affected by the inherent variability of the measurement system. A less variable alternative is the estimate of potency of a test product relative to a reference preparation. Dinse and Umbach (2013) define the relative potency (RP) as the "ratio of equi-effective doses (reference divided by test)". In the log-scale, this ratio is the horizontal distance between the two concentration-response curves. To estimate the RP as a unique value, this distance must be constant through the entire dilution profile. Therefore, both the European Pharmacopeia (Ph.Eur) and United States Pharmacopeia (USP) require the parallelism between the test and reference preparation concentration-response functions be demonstrated. This parallelism is called statistical similarity. 4PL curves are parallel if they share lower asymptotes, upper asymptotes, and steepness. Statistical assessment of the curves, referred to as parallelism tests or similarity tests, are performed to judge parallelism.

Several authors have proposed equivalence testing to assess similarity, which requires the definition of acceptance limits. The USP chapter <1032> proposes derivation of these limits by repeated comparison of the reference to itself. This approach so controls the risk of rejecting parallelism in case of true parallelism (lab risk) but fails to account for the risk of accepting curves with non-similarity that may result in highly inaccurate RP estimation (consumer risk). Additionally, tests that separately assess the equivalence of each non-C_{50} parameter ignore their correlation, leading to poorly defined limits.

To address these challenges, we first propose a three-step derivation of acceptance limits which can be used for any similarity test. The first step generates the posterior distribution of the test statistics under true parallelism, using a Markov chain Monte Carlo method. The second step defines the ac-

ceptance limits by simulating reference-to-reference comparisons. The third step controls the error in RP estimation for curves that pass the parallelism test. In order to test the equivalence of each parameter simultaneously, an ellipsoid-shaped zone of declared similarity is proposed. This ellipsoid test is compared to several published tests and exhibits overall better performance.

The USP also recommends screening of data for outliers prior to similarity assessment and RP estimation. Existing and proposed outlier tests are compared, and robust regressions are shown to have better detection rates for 4PL curves than the tests proposed in the USP chapter <1010>.

Finally, a simple yet efficient and robust method to choose the concentration support points for the concentration-response curve is proposed.

Résumé

Les essais de puissance mesurent l'activité biologique ou la capacité de liaison d'une substance. Cette activité est mesurée du début à la fin du procédé de développement et production de produits pharmaceutiques, avec des objectifs différents à chaque étape. Une mesure populaire de la puissance est le C_{50}, qui représente la concentration qui provoque une réponse égale à 50% de l'intervalle de réponses. Le C_{50} est estimé en modelant la réponse d'une substance en fonction de sa concentration, souvent au moyen d'une fonction logistique à quatre paramètres (4PL).

L'estimation du C_{50} peut être affectée par la variabilité inhérente au système de mesure. Une alternative plus précise est de mesurer la puissance d'un produit testé relative à celle d'une préparation de référence. Dinse and Umbach (2013) définissent la puissance relative (PR) comme le « ratio de doses équi-effectives ». A l'échelle logarithmique, ce ratio est la distance horizontale entre les deux courbes concentration-réponse. Pour que la RP soit une valeur unique, cette distance doit être constante sur l'ensemble du profil de dilution. Dès lors, les pharmacopées européenne (Ph.Eur) et américaine (USP) requièrent que le parallélisme entre les fonctions concentration-réponse de la courbe test et la courbe de référence soit démontré. Ce parallélisme est appelé similarité statistique. Deux courbes 4PL sont parallèles si elles partagent les mêmes asymptotes hautes, asymptotes basses, et raideurs. Des tests statistiques, appelés tests de parallélisme ou tests de similarité, sont utilisés pour évaluer si les courbes sont suffisamment similaires.

Plusieurs auteurs ont proposé des tests d'équivalence pour évaluer la similarité, lesquels nécessitent de dériver des limites d'acceptation. Le chapitre <1032> de l'USP propose de déterminer ces limites en comparant la référence à elle-même un grand nombre de fois. Cette approche permet de contrôler le risque de rejeter des courbes en cas de vrai parallélisme (risque laboratoire), mais pas le risque d'accepter des courbes dont le non-parallélisme peut provoquer de larges erreurs d'estimation de la PR (risque client). De plus, les tests qui évaluent chaque paramètre hors-C_{50} séparément ignorent leur corrélation, ce qui mène à des limites mal définies.

Pour résoudre ces problèmes, nous proposons d'abord une méthode de définition des limites d'acceptation qui peut être utilisée pour n'importe quel test de similarité. La première étape de cette méthode génère la distribution

postérieure des statistiques de tests en cas de vrai parallélisme, usant des méthodes de Monte-Carlo par chaînes de Markov. La seconde étape définit les limites d'acceptation en simulant des comparaisons référence-référence. La troisième et dernière étape contrôle l'erreur d'estimation de la PR pour les courbes qui réussissent le test de parallélisme. Afin de tester simultanément l'équivalence de chaque paramètre, une zone de similarité à forme ellipsoïde est proposée. Ce test ellipsoïde est comparé à plusieurs tests publiés, et présente globalement de meilleures performances.

L'USP recommande aussi de dépister les valeurs aberrantes dans les données avant de tester le parallélisme et estimer la RP. Nous comparons des tests de détection de valeurs aberrantes et démontrons que les régressions robustes ont des meilleurs taux de détection pour les courbes 4PL que les tests proposés dans le chapitre <1010>.

Enfin, nous proposons une méthodologie simple, efficace et robuste pour choisir les concentrations utilisées pour modéliser la courbe concentration-réponse.

Acknowledgements

Cette thèse n'aurait pu voir le jour sans une multitude de collaborations avec de nombreuses personnes.

Je tiens tout d'abord à remercier les Docteurs Bruno Boulanger et Pierre Lebrun de m'avoir lancé sur ce projet et guidé tout au long de sa réalisation. Aussi, et peut-être surtout, je veux les remercier chaleureusement de m'avoir transmis leur passion pour les statistiques pharmaceutiques et m'avoir formé tout au long de ma carrière. Je ne serais pas le quart du statisticien que je suis sans leur aide plus que précieuse et leur serai éternellement reconnaissant.

Je remercie aussi le Professeur Anne-Françoise Donneau d'avoir accepté le rôle de promotrice de thèse malgré un sujet hors de sa principale zone d'intérêt académique, ainsi que pour sa guidance et pour son soutien continuel qui ont permis à ce travail de progresser et d'aboutir.

Durant la réalisation de cette thèse, un important défi à relever fut de s'assurer que le lien entre statistiques appliquées et pharmacologie soit maintenu. Cela n'aurait pas été possible sans la contribution du Docteur Julien Hanson, que je remercie sincèrement.

I also thank jury members Professor Helena Geys, Professor Fulvio De Santis, and Doctor Éric Ziemons for taking their personal time to read and review my work, and make insightful suggestions for improvements.

Le projet original est né au sein de la société Arlenda et n'aurait été réalisable sans un soutien logistique de l'ensemble de l'équipe. Merci au Docteur Benoit Verjans de m'avoir permis d'utiliser à volonté les ressources de l'entreprise pour la réalisation de mon travail. Outre les Docteurs Bruno Boulanger et Pierre Lebrun, déjà cités plus haut, de nombreux collègues chez Arlenda ont joué des rôles directs ou indirects dans l'accomplissement du projet. Ne pouvant tous les nommer ici, je remercie en particulier les Docteurs Éric Rozet et Réjane Rousseau de m'avoir initié au monde des statistiques pour potency assays.

I would like to also offer my warmest thanks to my Arlenda US-based colleagues Tara Scherder and Katherine Giacoletti. Their precious advice and coaching have made me a better statistician in many ways and have particularly improved my ability to communicate statistics to non-statisticians.

In addition to my Arlenda colleagues, I was extremely lucky to work with and be mentored by many brilliant statisticians working on the field of potency assays, through my collaboration with MedImmune and the United States Pharmacopeia. I thank Doctors David Lansky, Harry Yang, Steven Novick, Bill Pikounis, and Mister Tim Schofield, for their valuable help. Many times have their experience helped me make progress during my research. Thanks also to Doctors Lingmin Zeng and Binbing Yu for being co-authors on papers used in this thesis, and Miss Sarah Janssen for her help in implementing parts of this work.

Ayant rapidement réalisé que je n'étais pas fait pour les biostatistiques cliniques, j'aurais probablement changé d'orientation de carrière sans la guidance du Professeur Bernadette Govaerts, je la remercie sincèrement. Merci aussi aux Professeurs Gentiane Hasbroeck et Philippe Lambert d'avoir facilité mon transfert vers la Faculté de Médecine.

Il m'est impossible de remercier personnellement tous les amis qui m'ont soutenu durant mes études de deuxième et troisième cycles, rencontrés lors de mes études secondaires et supérieures, ou au sein de la Royale Union Liégeoise des Étudiants de Louvain-La-Neuve et l'Ordre du Centaure. Merci à tous ceux qui se reconnaissent dans ce paragraphe pour les moments de détentes plus que nécessaires. En particulier, je voudrais remercier Raphaël Joie, dont l'amitié n'a de cesse d'augmenter ma soif d'accomplir; Lénaïc Damman, qui m'a montré que s'il a été capable d'obtenir un diplôme avec un travail à plein temps, une femme et deux enfants, je n'avais aucune excuse pour ne pas finaliser ma thèse; Sabrina El Bachiri, qui m'a aidé à réaliser qu'on peut devenir statisticien sans être trop obsédé par les statistiques; et Micael de Abreu Caldas, qui m'a convaincu d'abandonner le cursus universitaire qui ne m'allait pas du tout et sans qui je n'aurais pas eu le courage de démarrer mes études en statistiques. Thank you also to my friend Doctor Angel Lu, who shared with me the struggle of managing a thesis while working at Arlenda.

Merci à ma famille, et en particulier à mes parents, de m'avoir continuellement soutenu durant mon parcours universitaire.

Finally, I would like to thank my companion Elizabeth Scherder, for her love, for her patience, for bringing joy into my life and overall for making it easier for me to be me.

Contents

Chapter 1

Introduction

1.1 Potency Assays

An assay is the determination of the activity, potency, or concentration of a substance (also called the analyte) [1]. Potency assays measures the concentration or amount needed to produce a defined effect [2]. They can also be used to measure the biological activity or binding affinity of a sample product (or lot of product) relative to a reference preparation. For example, in the manufacture of a biotherapeutic or vaccine, the potency of material from a new batch may be measured relative to the potency of reference batch material [3]. A common way to estimate a potency is through the use of serial dilution assays, in which the response is measured at several concentrations providing a concentration-response (or dilution-response) function [4]. These serial dilutions are commonly performed within a 96-well plate (8 rows and 12 columns, see Figure 1.1). Due to laboratory constraints, one series of dilution is usually performed within one row or one column, although it is less than ideal because location effects within one plate may be present [5]. Whenever possible, the vials on the border of the plates are not used or used for control samples to avoid lowering the assay performance due to an edge effect [6].

Two main types of potency assays are encountered in pharmaceutical development: ligand binding assays and biological assays or bioassays.

Source: https://www.chromtech.com/

Figure 1.1: Representation of a 96-well plate.

1.1.1 Ligand Binding Assays

Findlay and Khan [7] define ligand-binding assays (LBA) as assays "in which the key step is an equilibrium reaction between the ligand (analyte) and a binding molecule, most often a protein and, in many cases, a specific antibody or receptor directed against the ligand of interest". In other words, they quantify the strength of the interaction between two molecules [8].

Some drugs are developed to mimic the action of the endogenous transmitter by binding to a receptor to induce a biological response, acting like agonists, or without inducing the response, acting like antagonists [9]. LBAs are generally limited as a drug-screening technique, because measuring the binding affinity of the analyte for a target receptor does not provide any information about its effects on the body while it is off the target [10]. Additionally, in most cases, there is little relationship between the strength of binding of a ligand to a site and its actual potency [11]. However, if the desired effect of a drug is only due to the binding, LBAs can be used to mea-

sure potency [12]. This is for example the case for monoclonal antibodies, which are used for the treatment of a variety of conditions, including but not limited to cancer [13, 14], autoimmune diseases [15, 16], inflammation [17], and infectious diseases [18]. Monoclonal antibodies are also used in vaccine development to understand the immune response to injected antigens [19–21].

The most common LBA used to measure the potency of monoclonal antibodies is the enzyme-linked immunosorbent assay (ELISA) [22–25]. During an ELISA, the substance of interest is bound to a solid surface (i.e. the surface of a well in a 96-well plate), and the strength of the binding reaction is quantified by measuring the optical density on each vial. ELISA assays are also very commonly used for the diagnostic of food allergies [26, 27] and multiple diseases, such as HIV [28], Lyme [29], rotavirus [30], dengue [31], hepatitis [32], and many others.

1.1.2 Biological Potency Assays

Bioassays, also called cell-based assays, measure the concentration, efficacy, and potency of a substance by assessing the effect it produces in living matter [33, 34]. They tend to be more complex than LBAs and the results are less precise, because the metabolic state of living cells varies from day to day [35]. Despite their disadvantages, they are sometimes necessary. Live vaccines for example, that contain attenuated viruses or bacteria (e.g. the measles and the polio vaccines), require bioassays to ensure successful infectivity. Common responses in *in vitro* bioassays for vaccines are the number of infected, lysed, or bound cells [36, 37]. *In vivo* assays, performed on animal, also exist. These are however sometimes highly variable and time consuming [38], and also pose some ethical concerns that have led the European Union to issue a directive to replace, reduce and refine animal assays [39, 40]. Biological assays can be direct — the concentration of both reference and test preparations are directly measured — or indirect — the ratio of concentrations from a reference and test preparation is measured, and the response can can be either continuous [41] or binary [42]. This work focuses on indirect potency bioassays with continuous responses in the context of therapeutic proteins (biologics) and vaccine development.

1.1.3 Applications

ELISA assays can be used in areas unrelated to drug products, such as growth-rate studies in botanical sciences [43] and detection of dioxin or dioxin-like compounds in wastes and the environment [44, 45]. This dissertation focuses on biotherapeutics and vaccine applications.

Bio and binding assays are used thorough the entire pharmaceutical product process. In the drug discovery phase, potency evaluation is mostly performed to rank and select valuable candidate molecules. During the development of a biologic or, sometimes, a small molecule, the safety and the efficacy evaluations can be affected by an immune response to the product. In pre-clinical studies, potency assays are used to quantify the effect of the protein and help determine the therapeutic dose for the intended purpose of the drug [46]. Through the clinical development, they are used to characterize the product, and develop and optimize manufacturing process [47]. For vaccines and biologics, they are also used in the later phases of clinical trial, as well as biological license applications and routine manufacturing, for lot release testing and stability assessment [48, 49].

1.2 Potency Determination

1.2.1 Concentration-Response Functions

A common measure of the potency of a substance is the C_{50}: concentration that yields a response equal to 50% of the range between the baseline (response when the concentration tends to 0) and the maximum (response when the concentration tends to infinity). It is estimated by modeling the substance response as a function of its concentration.

Concentration-response functions in potency assays tend to be sigmoid (see Figure 1.2). Below a certain concentration, there is no interaction between the substance of interest and the live cells or binding molecule. When the concentration increases, so does the effect of the substance, up to a certain point where no more effect can be produced by increasing the dose. The most used model to describe serial-dilution assays is the four-parameter logistic (4PL), and a common parametrization of this model was proposed by

Rodbard and Hutt [50]:

$$y_{ij} = yMax_i + \frac{yMin_i - yMax_i}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^{S_i}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{1.1}$$

where:

- y_{ij} is the response at concentration x_{ij} from curve i;
- ϵ_{ij} is the measurement error;
- $i = R, T$ correspond respectively to the reference and test preparations;
- $j = 1, \ldots, n$ corresponding to each of the n observations per curve;
- σ^2 is the measurement variability;
- $yMin_i$ and $yMax_i$ are respectively the lower and the upper asymptotes of curve i;
- S_i is a measure of the steepness of curve i;
- C_{50} is the concentration at the inflection point of curve i. This is the concentration needed to reach half of the response range between $yMin_i$ and $yMax_i$.

While mathematically equivalent, the following parametrization presents computational advantages [51]:

$$y_{ij} = yMax_i + \frac{yMin_i - yMax_i}{1 + \exp\left(S_i\left(\log(x_{ij}) + c_i\right)\right)} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{1.2}$$

where $c_i = \log(C_{50_i})$. We note $\theta_i = [yMin_i, yMax_i, c_i, S_i]^t$.

1.2.2 Relative Potency Calculation

Due to the inherent variability in test systems, using C_{50} as the measure of potency is not always possible [47]. A less variable option is to measure the potency relative to a reference preparation. Dinse and Umbach (2013) define the relative potency (RP) in serial dilution assays as the "ratio of equi-effective doses (reference divided by test)" [52]. Graphically, in the log-scale, this ratio is the horizontal distance between the two concentration-response curves. For the RP to be a unique value, this distance must be constant through the entire dilution profile (see Figure 1.3). Therefore, a necessary step before calculating the RP is to demonstrate the parallelism between the test and reference preparation concentration-response functions [47,53]. This parallelism is called statistical similarity.

1.2.3 Parallel Curves and Parallel Lines

4PL curves are a simple horizontal shift from one another in the log scale if and only if they share the same lower asymptotes, upper asymptotes and steepness's . In case parallelism is declared, the RP can be calculated by the ratio of C_{50}s after fitting the model :

$$y_{ij} = yMax + \frac{yMin - yMax}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^{S}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{1.3}$$

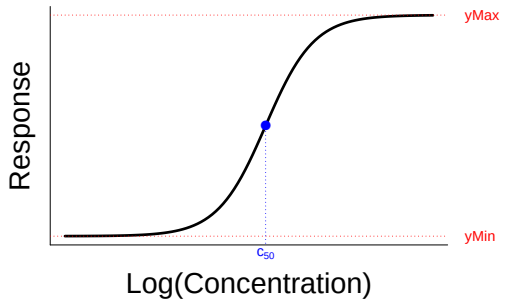


Figure 1.2: Concentration-response sigmoid curve

This model is an adaptation of equation 1.1 with common $yMin$, $yMax$ and S for both curves.

Historically, however, fitting nonlinear models was not easily done. Before the rise of user-friendly statistical software, the C_{50} had to be approximated from the serial dilution results using simple methods such as Spearman-Karber and Reed-Muench [54, 55]. These methods are still used in quantal assays [56]. Another common way to address this issue was to 'cut' the asymptotes from the model and fit y as a linear function of the remaining portion of $\log(x)$ (see Figure 1.4). If parallelism between the two lines is demonstrated, the following model is fitted:

$$y_{ij} = \beta_{0_i} + \beta_1 \times \log(x) + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{1.4}$$

where:

- β_{0_i} is the intercept of line i;
- β_1 is the slope, common for both curves.

The RP in this case is calculated by:

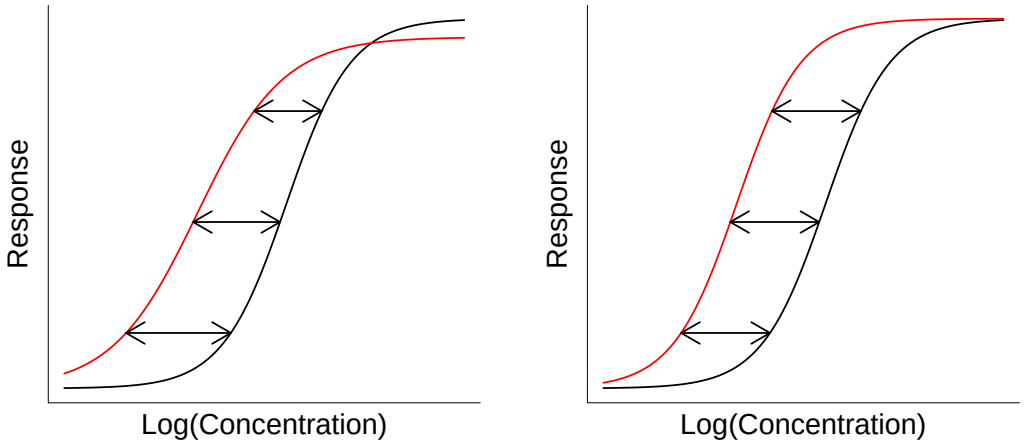


Figure 1.3: Parallel (right) VS non-parallel (left) curves and their effect on horizontal distance consistency

$$RP = \exp\left(\frac{\beta_{0_T} - \beta_{0_R}}{\beta_1}\right) \tag{1.5}$$

Log-linear approximation of sigmoid curves is not advised, because the obtained relative potency estimate is less accurate than when 4PL curves are used [47, 57, 58], and is usually not necessary. However, some assay designs may not allow to observe the full dilution profile. Parallel line assays are therefore still used nowadays in some cases [59].

1.2.4 Statistics for Nonlinear Models

Linear models present computational advantages. The maximum likelihood estimation of the model parameters as well as their respective standard errors can be directly calculated using matrix algebra. This is not the case for nonlinear models, such as the 4PL, and specific software are required to estimate the parameters.

Nonlinear models take the form:

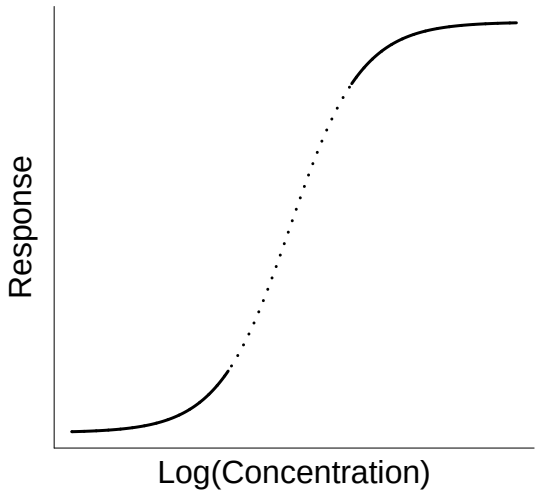


Figure 1.4: Line and Curve Assays

$$y = f(x, \theta) + \epsilon \tag{1.6}$$

where:

- x is a vector of independent variables;
- y is a vector of a dependent variable;
- ϵ is the vector of errors;
- θ is the vector of parameters;
- f is a nonlinear function.

For example, in the 4PL as described in Equation 1.3,

$$\theta = [yMin, yMax, C_{50_R}, C_{50_T}, S]^t$$

and

$$f(x, \theta) = yMax + \frac{yMin - yMax}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^S}$$

Nonlinear models can be fit on data using classic frequentist or Bayesian methods.

1.2.4.1 Using Frequentist Statistics

Fitting a nonlinear model requires to use a numerical method to find the combination of parameters that minimizes the residual sum of squared errors (RSSE) [60, 61]. The variance-covariance matrix of the obtained parameter estimates is then approximated using Taylor expansions [62]. This approximation is decent with large sets of data but may be inaccurate with small sample sizes [63,64]. Due to their cost and complexity, potency assays usually belong to the second category, and standard errors of 4PL curve parameters tend to be poorly estimated. Frequentist methods are therefore inappropriate to make any inference on model parameters. However, they are convenient to provide the maximum likelihood parameter estimates, noted $\hat{\theta}$, including the RP, and test for parallelism.

Due to their computational complexity, these methods used to be unavailable for the general public. Nowadays, many user-friendly software packages fit 4PL curves using numerical methods. Some examples are JMP [65], SoftMax Pro [66], PLA [67] and GraphPad Prism [68]. Scripting software also provide solutions for fitting 4PL easily, such as the R package *DRC* [69] or the SAS procedure NLMIXED [70].

1.2.4.2 Using Bayesian Statistics

Another advantage of linear models is that, if the necessary assumptions are respected, the distribution of each model parameter estimate is known. Closed-form solution (or approximated solutions in the case of hierarchical models) for parameter inference and future observation predictions are therefore directly available. In nonlinear models, the distribution of the parameters is known only if the sample size approaches infinity [71]. Bayesian statistics offer an alternative. Rather than considering the unknown parameters as a constant value, they are treated as random variables about which we have a certain degree of prior knowledge [72].

From equation 1.6, the Bayes theorem gives:

$$pdf(\theta|data) \propto pdf(data|\theta) \times pdf(\theta) \tag{1.7}$$

where:

- *pdf* is the probability density function;
- $pdf(\theta)$ is the prior distribution of θ, representing the prior knowledge;
- $pdf(data|\theta)$ is the likelihood of observing the data knowing θ, and the function that is maximized in frequentist statistics;
- $pdf(\theta|data)$ is the posterior distribution of θ.

An obvious advantage of Bayesian statistics is the inclusion of $pdf(\theta)$. In frequentist statistics, only the observed data can be used to estimate and infer on parameters and make predictions. Including prior knowledge in the analysis increases the degrees of freedoms and therefore lower the necessary

sample size necessary to obtain the same level of uncertainty in the results. If no prior knowledge is available, non-informative distributions can be used for the prior, such as locally uniform or distributions with very large variances. Another advantage is that the only distribution that needs to be assumed is the one of ϵ (e.g. $\mathcal{N}(0, \sigma^2)$). Software packages mentioned in Section 1.2.4.1 each assume that $\hat{\theta}$ asymptotically follows a normal distribution. Therefore, when available, the provided confidence intervals are symmetrical and centered in $\hat{\theta}$. This assumption is often incorrect. Bayesian statistics do not involve asymptotic theory and exact posterior distributions are obtained.

A drawback of Bayesian methods is that they are not as easily accessible in user-friendly software. Closed-form solutions are not available for nonlinear models and distributions of each model parameter are sampled from using Markov Chain Monte Carlo methods (MCMC). The most common MCMC algorithms for nonlinear models are Metropolis-Hasting [73] and Hamiltonian (or Hybrid) Monte Carlo [74]. Software to use such algorithms to fit 4PL curves include Stan [75], JAGS [76] and SAS procedure MCMC [77].

An example of Bayesian methodologies applied to serial dilution assays is presented in Appendix A

1.2.5 Similarity Testing for Potency Assays

Parallelism is achieved if $yMax_R = yMax_T$ and $yMin_R = yMin_T$ and $S_R = S_T$ or, for the linear approximation, $\beta_{1_R} = \beta_{1_T}$. Because $\sigma^2 > 0$, perfect parallelism is never observed. Instead, statistical tests are performed to assess if the curves are parallel enough. These tests are referred to as parallelism tests or similarity tests.

1.2.5.1 Difference Tests for Parallel Curve Assays

Difference tests rely on $p - values$. The hypotheses are defined as follow:

- $H_0 : f_R(x, \theta) = f_T(x \times RP, \theta)$
- $H_a : f_R(x, \theta) \neq f_T(x \times RP, \theta)$

The first test that was proposed to assess similarity was a lack-of-fit $F-Ratio$ [78]. This test compared the RSSE due to non-parallelism to its RSSE due to other reasons. It is realized by fitting both Equation 1.1 and 1.3 to the same pair of curves (see Figure 1.5), then calculate the test statistics:

$$F_{nonpar} = \frac{(RSSE_{constrained} - RSSE_{full})\,(df_{constrained} - df_{full})}{(RSSE_{full})\,(df_{full})} \tag{1.8}$$

where $RSSE_{full}$ and $RSSE_{constrained}$ are the RSSE of the models obtained from Equation 1.1 and 1.3, respectively, and df_{full} and $df_{constrained}$ are their associated degrees of freedom. Under H_0, $F_{nonpar} \sim F_{(df_{constrained}-df_{full}),df_{full}}$.

This test statistics is known to be overly sensitive to small shift from parallelism when σ^2 is small, and incapable of detecting drifts when σ^2 is high. Gottschalk and Dunn (2005) proposed an alternative approach based on the difference between the weighted RSSEs (wRSSE) of the model obtained from Equations 1.1 and 1.3 [79]:

$$RSSE_{nonpar} = wRSSE_{constrained} - wRSSE_{full} \tag{1.9}$$

Under H_0, $RSSE_{nonpar} \sim \chi^2_{(df_{constrained}-df_{full})}$ [79].

A drawback of this χ^2-test is that it requires weighted regressions, which is not recommended for serial dilution models [80, 81]. Assuming that a weighting is actually needed, the correct weights are unknown and tend to be poorly estimated. Additionally, both tests declare similarity in case of lack of statistical significance to demonstrate non-parallelism. This is fundamentally flawed, as failure to reject the null hypothesis does not mean that it is true.

1.2.5.2 Equivalence Tests for Parallel Curve Assays

Equivalence tests revert hypotheses such that statistical significance is required to demonstrate parallelism instead of the other way around [82]. Jonkman and Sidik (2009) proposed to test the equivalence of each curve parameters, except the C_{50}s, using an intersection-union test [83]. Yang et al. (2012) proposed the following hypotheses as an alternative [84, 85]:

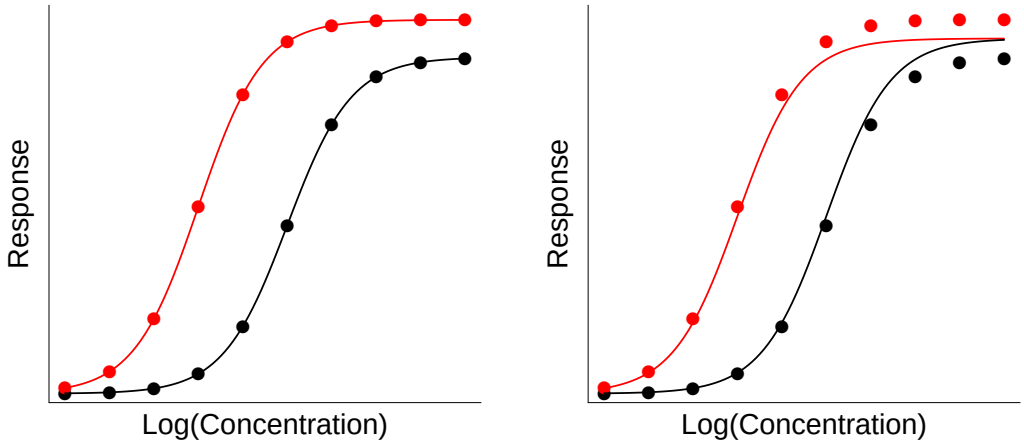


Figure 1.5: Full and constrained 4PL models

- $H_0 : r_1 \leq D_{L1}$ or $r_1 \geq D_{U1}$ or $r_2 \leq D_{L2}$ or $r_2 \geq D_{U2}$ or $r_3 \leq D_{L3}$ or $r_3 \geq D_{U3}$
- $H_a : D_{L1} < r_1 < D_{U1}$ and $D_{L2} < r_2 < D_{U2}$ and $D_{L3} < r_3 < D_{U3}$

with

$$r_1 = \frac{yMax_T}{yMax_R}$$
$$r_2 = \frac{yMax_T - yMin_T}{yMax_R - yMin_R}$$
$$r_3 = \frac{(yMax_T - yMin_T)S_T}{(yMax_R - yMin_R)S_R}$$

and where D_{Lk} and D_{Uk} are respectively the lower and upper equivalence limits for $r_k, k = 1, 2, 3$.

The reason for r_2 is that $yMin$ is often close to 0, so very small derivation in lower asymptotes would give very high variation in $\frac{yMin_T}{yMin_R}$. r_3 is the ratio of slopes at the inflection point, and its estimator is less variable than the ratio $\frac{S_T}{S_R}$. Through this dissertation, this test is referred to as 'Hyper-Rectangle'

(HR) test, because the three-dimensional space between all D_{Lk} and D_{Uk} forms a hyper-rectangle. We work with the log-ratios as suggested by Berger and Hsu (1996) when working with ratios [86].

Lansky (2019) proposed to work with differences, rather than ratios, scaled on the long-term average of the reference curve parameters [87] . Assuming that the long-term average is a good estimator of the true parameter and can be treated as a constant, it would provide the computational advantages of working with additive rather than multiplicative interactions. This test was not yet published. Therefore we may not have all the information necessary to replicate it as is, and we also never had the opportunity to try it on non-simulated data. For these reasons, we did not include it in the comparisons performed through this dissertation.

Novick et al. (2012) claim that analyzing ratios of parameters is not enough to declare similarity between the entire nonlinear profiles [88, 89]. Instead, they suggest looking at the maximum departure between the confidence intervals of the reference curve fit and the test curve fit after the optimal horizontal shift (the relative potency), within a certain range of concentration $[x_L, x_U]$. They first used a Bayesian method to calculate the confidence intervals [88] and later proposed a frequentist approximation [90]. The hypotheses for this test are

- $H_0 : \min\limits_{RP} \max\limits_{x_L, x_U} |f(x, \theta_R) - f(x \times RP, \theta_T)| \geq \delta$
- $H_a : \min\limits_{RP} \max\limits_{x_L, x_U} |f(x, \theta_R) - f(x \times RP, \theta_T)| < \delta$

where δ is the maximum accepted departure from parallelism to declare similarity.

1.2.5.3 Tests for Parallel Line Assays

Two straight lines are parallel if they share the same slopes. Difference tests based on composite measures such as F_{nonpar} and $RSSE_{nonpar}$ would be applied by comparing the fit of both lines separately against a model with common slopes. An equivalence test for the slopes is:

- $H_0 : \frac{\beta_{1_T}}{\beta_{1_R}} \leq D_{L\beta_1}$ or $\frac{\beta_{1_T}}{\beta_{1_R}} \geq D_{U\beta_1}$

- $H_a : D_{L\beta_1} < \frac{\beta_{1_T}}{\beta_{1_R}} < D_{U\beta_1}$

The confidence interval for the ratio of the slopes can be either obtained directly from the posterior distributions of β_{1_T} and β_{1_R} if the models are fit using Bayesian methods. Otherwise, frequentist approximation exists, such as the Fieller's theorem [91, 92].

1.3 Challenges and Opportunities for Improvement

Historically, similarity assessment was only performed from late clinical phases forward, as the ability for a substance to create a biological or binding reaction rather is of more interest than the precise estimation of potency in the discovery phases. However, since the publication of the Quality by Design (QbD) concepts in ICH-Q8 in 2009 [93] for the development and validation of processes, there has been an increasingly convergent agreement that the same paradigm also applies to LBAs and bioassays. Several publications introduced the lifecycle management concept of analytical methods, an approach closely related to QbD [94–97]. Indeed, both QbD and lifecycle management processes start with the identification of the objectives and requirements, then knowledge building during method development, validation/qualification, and finally, development of a control strategy to permit a continued improvement [98]. Conceptually, this means that the validation and routine use of the assay should be considered important from the moment it is developed, and similarity measures should always be considered as critical quality attributes.

While the literature now wildly advises against the use of difference tests for parallelism, equivalence present other challenges. The principal challenge is that equivalence tests require equivalence margins to define what constitutes an unimportant difference. USP <1032> (2012) [47] proposes four methods to derive those margins. The first three methods use historical data to estimate variability of comparing the reference to itself, while the fourth suggests including a risk analysis regarding the implications of the chosen equivalence margins on the quality of the relative potency estimation. Using historical data, when available, allows to control for the risk of rejecting

similarity in case of true parallelism (lab risk) by accounting for expected variability when comparing the reference to itself; however, using historical data alone fails to account for the risk of rejecting curves which result in high error in relative potency estimation (consumer risk). In addition, historical data are not always available or are limited in early development stages of the assay. It is always possible for laboratories to generate reference to reference comparison, but this is costly both in time and money. Another challenge, specific to the HR test, is that is does not account for the correlation of the ratios of curve parameter estimates under true parallelism. Marginal equivalence limits are not optimal for such correlated test statistics.

The USP<1032> guidelines recommend the screening of potency assay data for outliers prior to performing a RP analysis. The guidelines, however, do not offer advice on the size or type of outlier that should be removed prior to model fitting and calculation of RP. Computer simulation was used to investigate the consequences of ignoring the USP<1032> guidance to remove outliers. For biotherapeutics and vaccines, outliers in potency data may result in the false acceptance/rejection of a bad/good lot of drug product. One or more outliers in the concentration-response data, may result in a failure to declare similarity or may yield a biased RP estimate. Our findings generally support the USP<1032>. The outlier tests proposed by the USP<1010>, however, are not well suited for concentration-response curves analysis, and many outliers remain undetected, and more suitable tests are needed.

Another challenging aspect of potency assays is choosing the ideal concentrations for the concentration-response curve analysis. Common optimal design methods are not suited for this type of analysis, as they fail to account for the constraint in a laboratory as well as the between-run variabilities that affect the estimation of the relative potency and the non-similarity test statistics.

Finally, every time a parallelism test is proposed, this test is claimed to be the correct way to address parallelism, and their advantages are presented using limited examples. Assay scientists are left with statistical papers that do not provide a real indication on which test to use in which situation.

1.4 Structure of the manuscript

In Chapter 2, we propose a three-step derivation of the zone of declared similarity to control for both consumer and lab risk, even when historical data are not available. In addition, we propose a novel way to test for parallelism that accounts for the test statistics correlations.

By definition, hypothesis testing is used when a sample of observations is assumed to be representative of an entire population. In many cases, however, the primary interest is in the relative potency estimate, which is affected by similarity of the samples themselves. Similarity is then a necessary condition for each sample, to ensure that it is suitable to estimate the RP. Each sample is, in that case, it's own population, because the accuracy of the RP will depend on the estimated ratio of parameters, not the unknown true value. Equivalence tests based on confidence intervals are therefore not necessary for sample suitability, as confidence intervals are used to determine a range of likely values for the unknown true value. Through this dissertation, we present similarity tests only with the necessary condition that two observed curves must meet to be declared parallel, rather than with null and alternative hypothesis.

In Chapter 3, concentration-response curves for test and reference were computer generated with constant RP from four-parameter logistic curves. Single outlier, concentration outlier, and whole-curve outlier scenarios were explored for their effects on the similarity testing and on the RP estimation. Though the simulations point to situations for which outlier removal is unnecessary, the results generally support the USP<1032> recommendation and illustrate the impact on the RP calculation when application of outlier removal procedures are discounted. In Chapter 4, several outlier detection methods, including those proposed by the USP<1010>, are evaluated and compared through computer simulation. Two novel outlier detection methods are also proposed. The effects of outlier removal on similarity testing and estimation of relative potency were evaluated, resulting in recommendations for best practice.

In Chapter 5, we propose a way to find an efficient concentration range, easily calculated from an estimation of the curve parameters.

In Chapter 6, we compare the performances of the parallelism tests discussed in Section 1.2.5, as well as the novel test proposed in Chapter 2. This

extensive comparison evaluates scenarios from an optimal design (as defined in Chapter 5) without any issues, to a poor design with removal of outliers or increasing measurement variability.

Finally, in Chapter 7, we propose a general discussion on the work that was realized. We also present suggestions for follow-up research.

Chapter 2

Risk-Based Similarity Testing for Potency Assays Using MCMC Simulations

This Chapter is based on the article titled "A risk-based multivariate similarity test for potency assays", accepted for publication in *Statistics in Biopharmaceutical Research*.

2.1 Introduction

In relative potency (RP) assays, the RP can be determined by computing the horizontal difference between the log(concentration)-response functions of the test and reference [47], should they be linear or a nonlinear function. In both cases, the computed value is meaningful only if the concentration-response functions of the two products (or batches, lots, samples), are similar (or parallel). Indeed, similarity indicates that the biological activity, or binding affinity, within the assays is similar for both the test and the reference products [34, 79, 99]. That is, in the log(concentration) scale, the function of the test product is a horizontal shift from the reference standard's function [84]. If the functions are not parallel, the horizontal difference is not a single constant value over the concentration range modeled (see Figure 1.3). From a regulatory perspective, both the United States Pharmacopeia (USP)

and the European Pharmacopeia (EP) require that parallelism is assessed before the computation of relative potency [47, 53].

To model the concentration-response function of a preparation, literature commonly suggests the use of a four-parameter logistic (4PL) model. The most common of these was proposed by Rodbard and Hutt [50], described in Section 1.2.1:

$$y_{ij} = yMax_i + \frac{yMin_i - yMax_i}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^{S_i}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{2.1}$$

Between-plate or between-run variabilities, as well as covariance structure within one plate or run, can be identified in addition to the residual variability. These longitudinal attributes can have an effect on one or several parameters of the 4PL model. Within a single run (or a plate), variability can exist between two samples of the same product. Additionally, the observed residual variability may not be homogeneous across the concentration range. These aspects should be addressed while modeling the potency assay data. However, for simplicity, these model selection details are not discussed extensively in this paper and the residual variability is assumed to be the same across the concentration range. More details on heteroscedastic systems are presented in the discussion section of this paper.

Parallelism is accepted if the lower asymptotes, upper asymptotes and growth rates are similar between the two curves (1.3 right). If parallelism is demonstrated, the RP is then the estimated horizontal distance between the two curves on a log scale after fitting the model

$$y_{ij} = yMax + \frac{yMin - yMax}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^{S}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{2.2}$$

Equation 2.2 is an adaptation of Equation 2.1 with common $yMin$, $yMax$ and S for both curves. The horizontal distance between the two curves is then the ratio of the C_{50}s.

Since the beginning of the 21st Century, several authors have proposed equivalence testing to assess similarity between two concentration-response

functions [82–84]. The use of equivalence tests is also recommended by the USP [47].

In Section 2.2, we present a common existing test for parallelism, and present two challenges that this approach presents. In Section 2.3, we suggest solutions for the challenges: an alternative test that accounts for the correlations among the ratios of curve parameter estimates, and a three-step derivation of the zone of declared similarity to control for both consumer and lab risk, applicable to any similarity test, even if historical data is not available. In Section 2.4, we present a case study and in Section 2.5, we discuss our observations and how to perform similarity testing as a sample suitability test, rather than a hypothesis test.

2.2 Common Parallelism Test

A popular equivalence test is proposed by Yang et al. (2012) [84]. It suggests to separately fit the 4PL model to both reference and test curves and then compare the estimated ratio of the upper asymptotes (r_1), the ranges between the asymptotes (r_2), and the slopes at the inflection point (r_3) to pre-defined equivalence margins.

$$r_1 = \frac{yMax_T}{yMax_R}$$
$$r_2 = \frac{yMax_T - yMin_T}{yMax_R - yMin_R}$$
$$r_3 = \frac{(yMax_T - yMin_T)S_T}{(yMax_R - yMin_R)S_R}$$

They consider the following hypotheses

- $H_0 : r_1 \leq D_{L1}$ or $r_1 \geq D_{U1}$ or $r_2 \leq D_{L2}$ or $r_2 \geq D_{U2}$ or $r_3 \leq D_{L3}$ or $r_3 \geq D_{U3}$
- $H_a : D_{L1} < r_1 < D_{U1}$ and $D_{L2} < r_2 < D_{U2}$ and $D_{L3} < r_3 < D_{U3}$

with D_{Lk} and D_{Uk}, respectively, the lower and upper equivalence limits for r_k. This is tested by comparing the confidence interval of the ratios to their

respective equivalence limits. We suggest working with the log(ratios) for symmetry, as suggested by Berger and Hsu (1996) [86]. We also recommend working with the point estimates of the ratios rather than their confidence intervals. The reasons for this are provided in Section 2.5. For this test, curves are accepted as parallel if

$$D_{L1} < \log(\hat{r}_1) < D_{U1} \text{ and } D_{L2} < \log(\hat{r}_2) < D_{U2} \text{ and } D_{L3} < \log(\hat{r}_3) < D_{U3} \tag{2.3}$$

Other tests for parallelism have been proposed and rely on composite measures rather than curve parameters [79, 88, 90]. Summarizing the entire dose-response profile in one measure is convenient and sometimes a good alternative to ratios of parameters. This Chapter focuses on testing the curve parameters ratios.

A first challenge with relying on ratios is that marginal equivalence limits are not optimal for such correlated parameters. Figure 2.1 shows a 3D representation of a hypothetical distribution of log(ratio) estimates in case of true parallelism, contained within marginal limits (a way to derive those limits is presented in Section 3). It shows that the 3-dimensional space contained within the marginal limits for r_1, r_2 and r_3 has a lot of 'empty space'. This means that curves within those empty spaces are not contained in the true distribution. Similarly, some parts of the distribution fall outside the marginal limits, and those curves would therefore be rejected by the hyper-rectangle test.

A second challenge is that equivalence tests require equivalence margins to define what constitutes an unimportant difference. USP <1032> proposes four methods to derive those margins. The first three methods use historical data to estimate variability of comparing the reference to itself, while the fourth suggests including a risk analysis regarding the implications of the chosen equivalence margins on the quality of the relative potency estimation [47]. Using historical data, when available, allows one to control for the risk of rejecting parallelism in case of true parallelism (lab risk) by accounting for expected variability when comparing the reference to itself; however, using historical data alone fails to account for the risk of rejecting curves which result in high error in relative potency estimation (consumer risk). In addition, historical data are not always available or are sometimes limited in early stages of the assay.

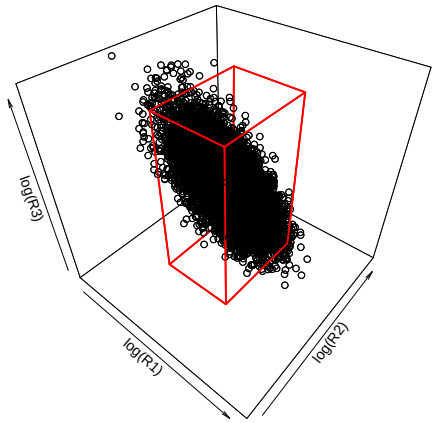


Figure 2.1: Distribution of curve parameter ratio estimates with marginal limits for equivalence test. The red hyper-rectangle represents the marginal margins for each log(ratio).

2.3 Proposed Solutions

In this section, we propose an alternative test, based on an ellipsoid in the three-dimensional space. We also suggest estimating the absolute relative error in relative potency estimation for curves that are declared similar, to ensure that the chosen zone of declared similarity does not allow curves which result in high error in relative potency estimation to pass similarity.

For the ellipsoid test, we assume that the estimate of the log(ratios) under true parallelism follow a three-dimension multivariate normal distribution centered in $[0, 0, 0]^t$ with variance-covariance matrix Σ. A point m falls within the ellipsoidal $100\,(1-\alpha)\,\%$ prediction region of a η-dimension multivariate normal distribution if

$$(m-\mu)^t\,\Sigma^{-1}\,(m-\mu) \leq q_{1-\alpha,\eta}$$

where $q_{1-\alpha,\eta}$ is the $(1-\alpha)^{th}$ quantile of a χ^2 distribution with η degrees of freedom [100]. A future pair of curves is considered parallel if it falls within the derived ellipsoid. We describe later in this section how to estimate $\hat{\Sigma}$ and derive an ellipsoid containing $100(1-\alpha)\%$ of parameter ratios for parallel curves. To test if a future pair of curve falls within the ellipsoid and therefore can be considered parallel, one therefore assesses if

$$\hat{\lambda}^t\hat{\Sigma}^{-1}\hat{\lambda} \leq q_{1-\alpha,3} \tag{2.4}$$

where $\hat{\lambda} = \log\left([\hat{r}_1, \hat{r}_2, \hat{r}_3]\right)$.

2.3.1 Step 1. Compute the posterior predictive distribution of each test statistic

From a qualification or validation set of plates, fit Equation 2.1 to the reference curves using a Bayesian platform such as Stan [75], JAGS [76], or PROC MCMC in SAS [77]. If historical data are available, they can be used to derive prior distributions. Like prior elicitation, model selection isn't addressed in this paper but is an important step, because between-run or between-plate variability often affects one or several parameters. Let $\Phi = p(\Theta|y, \mathcal{I})$ be the joint posterior distribution of the variance components and curve parameters

$\Theta = [yMin_R, yMax_R, C_{50_R}, S_R, \sigma^2]$, conditional on the observed data y and all prior information relevant to the analysis $\mathcal{I}$ (concentration range, number of concentration points, number of replicates per concentration, etc. as well as prior knowledge on any of the curve parameters or variance components).

Once Φ is obtained, one can generate thousands of runs containing one pair of reference curves each, both from the same draw of Φ and with the same $\mathcal{I}$ that will be used in routine analysis. From each generated pair of curves, one can fit Equation 2.1 and estimate the test statistics $\hat{r}_1$, $\hat{r}_2$, and $\hat{r}_3$. Doing this for all of Φ provides $\rho = p(\hat{r}|y, \mathcal{I}, \Theta)$, the posterior predictive distributions of future $\hat{r} = [\hat{r}_1, \hat{r}_2, \hat{r}_3]^t$ when two curves have truly equal θs.

2.3.2 Step 2. Derive zone of declared similarity while accounting for lab risk

In the case of a univariate test for parallelism, such as a slope ratio for parallel line assays, the proposed equivalence limits can be the $100(\alpha/2)^{th}$ and $100(1-\alpha/2)^{th}$ percentiles of the test statistic posterior predictive distribution, where α is the chosen lab risk. In the case of multiple independent test statistics, simple corrections like Bonferroni's [101], can be applied. Such simple correction cannot be applied in the case of ratios of curve parameters, because they are correlated. Instead, one can find the smallest three-dimensional space that contains $100(1-\alpha)\%$ of ρ by applying an optimization method such as Nelder and Mead's (1965) [61]. Note that the objective function of this optimization may not have a unique maximum. Therefore, we recommend testing multiple sets of initial values.

For the ellipsoid test, we have

$$\lambda = \log(\rho) \sim \mathcal{N}\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \hat{\Sigma}\right)$$

Where $\hat{\Sigma}$ is the estimated variance-covariance matrix of λ. Note that, because ρ is conditional on true similarity between two curves, the true center of λ is by definition $[0, 0, 0]^t$. It is important to draw a large number of effective samples from λ (say, 10,000) to have a good estimate of Σ.

2.3.3 Step 3. Control for consumer risks and take action if needed

The two steps above are an efficient alternative to the first three USP <1032> methods for equivalence margins derivation, because the MCMC simulation obviates the need to generate reference vs. reference comparisons via extensive experiments (although historical data, when available, are still useful to derive prior distributions). However, both the first three USP <1032> methods and the two steps above fail to consider the consumer risk. The fourth USP<1032> method suggests sensitivity analysis. This means calculating the consumer risk in terms of the relative error in relative potency (noted β) for curves that pass the parallelism test [47].

To do so, we suggest performing Monte Carlo simulations. From Φ, obtained in the first step, generate thousands of reference curves, with the same $\mathcal{I}$ that will be used in routine analysis. For each reference curve, generate a test curve from the same run, but with a deviation from true parallelism. For each of the simulations, randomly chose a shift in $yMin$, $yMax$, and S that respects the similarity limits of your chosen test, ensuring that the full range of ratios that would pass are contained within the simulation domain. Shifts in C_{50} can also be used to consider different true relative potencies, because while it will not affect the true ratios of the parameters, the horizontal shift may change how well those ratios are estimated.

For each pair of curves that passes the chosen parallelism test, fit Equation 2.2 and estimate the relative potency. Then calculate the absolute relative % error due to non-similarity:

$$\psi = 100 \times \left| \frac{\text{Observed RP} - \text{True RP}}{\text{True RP}} \right| \tag{2.5}$$

Repeating the operation many times provides $\Psi = p(\psi|y, \mathcal{I}, \Theta, \lambda)$, the posterior predictive distribution of ψ. Let β^* the 95^{th} percentile of Ψ. The acceptance limits, no matter the chosen test, should only be used if $\beta^* \leq \beta$, the maximum acceptable relative error. If $\beta^* > \beta$, actions to improve the assay should be undertaken. This will be considered further in the discussion of this paper.

2.4 Case Study

We blinded data from a qualification set of nine 96-well plates used to analyze the potency of a test vaccine lot relative to a reference lot. Each plate contains three replicates of the reference lot and three replicates of the test lot, across 10 concentration-points. The blinded data are presented in Figure 2.2.

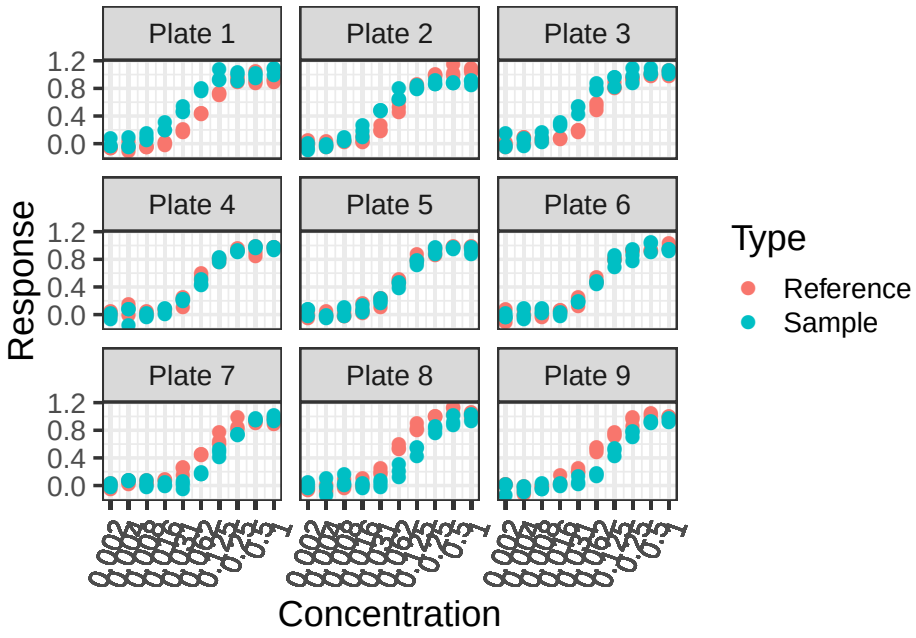


Figure 2.2: Presentation of each plate used in the case study.

2.4.1 Step 1. Compute the posterior predictive distribution of each test statistic

The following model was fit to the reference curves only:

$$y_{ijk} = yMax_k + \frac{yMin - yMax_k}{1 + \left(\frac{x_{jk}}{C_{50}}\right)^S} + \epsilon_{ijk} \tag{2.6}$$

where:

- y_{ijk} is the response of the i^{th} replicate of the reference product at the j^{th} concentration on the k^{th} plate;
- x_{jk} is the j^{th} concentration on the k^{th} plate;
- $yMin$, S and C_{50} are respectively the lower asymptote, steepness and C_{50}, common across plates;
- $yMax_k \sim \mathcal{N}\left(yMax, \sigma^2_{yMax}\right)$ is the upper asymptote of the k^{th} plate, with σ^2_{yMax} the between-plate variability of $yMax$;
- $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_e)$ is the error term, with σ^2_e the residual variability.

Equation 2.6 is a modification of Equation 2.2 where a between-plate variability affects the upper asymptote. Between-plate variability can affect one, several, or no model parameter and should be evaluated as part of model selection. Model selection isn't addressed in this dissertation. More details on Bayesian model selection can be found in Kruschke (2014) [102], and guidance on Bayesian methods for sigmoid curves can be found in Feng et al. (2011) [103] and Klauenberg et al (2011) [104].

The posterior distribution Φ of curve parameters was sampled by Hybrid MCMC using Stan. If no prior information is available, a flat, vague or vaguely-informative prior may be used for the parameters. We decided to use the flat prior so that the distributions were given by $\mathcal{U}(-\infty, \infty)$ for each curve parameter and $\mathcal{U}(0, \infty)$ for variance components. Improper priors for precision instead of variance is preferred in some statistical software. We however used Stan, which uses standard deviations, making it more intuitive to use improper prior distributions for variances. Vaguely (or weakly) informative prior distributions may also be used to aide MCMC convergence. In our particular case, however, convergence was not an issue with improper priors (see Figure 2.3). Results using informative prior distributions are presented in Section 2.4.4.

Two independent chains were run with each 300,000 draws, including a warm-up of 50,000 draws, thinning every 10 draws. This resulted in 25,000 posterior draws per chain and 50,000 posterior draws total. Each model parameter had an effective sample size of 45,000 or more, and the trace plots of the MCMC chains are presented in Figure 2.3.

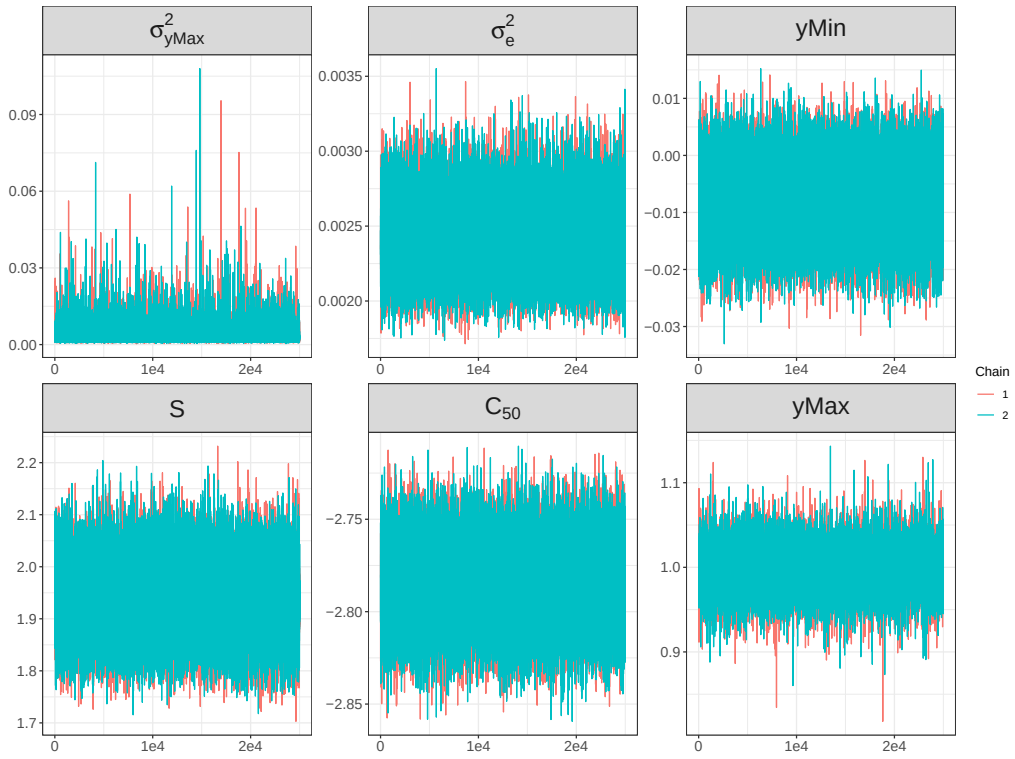


Figure 2.3: Trace plot of each model parameter when using non-informative prior distributions. From left to right: σ^2_{yMax}, σ^2_e, and $yMin$ on top; and S, C_{50}, and $yMax$ in the bottom

From the joint posterior distribution, we generated 50,000 plates with two reference curves each, with three replicates per curve, at each of the 10 concentration points used in the qualification set of plates. For each simulated plate, we fitted Equation 2.2 using ordinary least square and calculated $\hat{\lambda}$.

2.4.2 Step 2. Derive zone of declared similarity while accounting for lab risk

We wanted a lab risk of $\alpha = 1\%$. We calculated the limits for the hyper-rectangle from the 50,000 $\hat{\lambda}$s, using Nelder and Mead to ensure finding the actual smallest hyper-rectangle containing $(1 - \alpha)\,\%$ of the posterior predic-

tive distribution of $\hat{\lambda}$ in the three-dimensional space. The estimated limits for the hyper-rectangle test are presented in Table 2.1. Exactly 99% of the calculated $\hat{\lambda}$s respected Equation 2.3.

Table 2.1: Calculated limits for the hyper-rectangle test in natural log and original scale when using non-informative prior distributions

Ratio	**Limits in log scale**	**Limits in original scale**
r_1	[-0.104; 0.104]	[0.901; 1.110]
r_2	[-0.116; 0.116]	[0.890; 1.123]
r_3	[-0.377; 0.377]	[0.686; 1.459]

The variance covariance matrix $\hat{\Sigma}$ of the posterior predictive distribution of $\hat{\lambda}$ was also calculated:

$$\hat{\Sigma} = \begin{bmatrix} 0.0009 & 0.0011 & -0.0014 \\ 0.0011 & 0.0019 & -0.0024 \\ -0.0014 & -0.0024 & 0.0134 \end{bmatrix}$$

We observed that 98.7% of the calculated $\hat{\lambda}$s respected Equation 2.4.

Figure 2.4 shows a two-dimensional representation of the posterior predictive distribution of $\hat{\lambda}$. The blue points are inside the ellipsoid and the red-rectangles are the limits of the hyper rectangle. We observe that the corners of the rectangles are not contained in the ellipsoid.

The colored curves on the left panel of Figure 2.5 represent each of the eight corners of the hyper-rectangle. These curves are very dissimilar compared to the reference curve (black curve), but they are within the 99%-coverage hyper-rectangle and would therefore be accepted as parallel using that test. This issue confirms the claim that equivalence of parameters does not mean equivalence of distribution [88]. However, this is probably because equivalence tends to be evaluated marginally. The colored curves on the right panel of Figure 2.5 represent each of the 6 ellipsoid summits. These curves are visibly more similar to the reference curve.

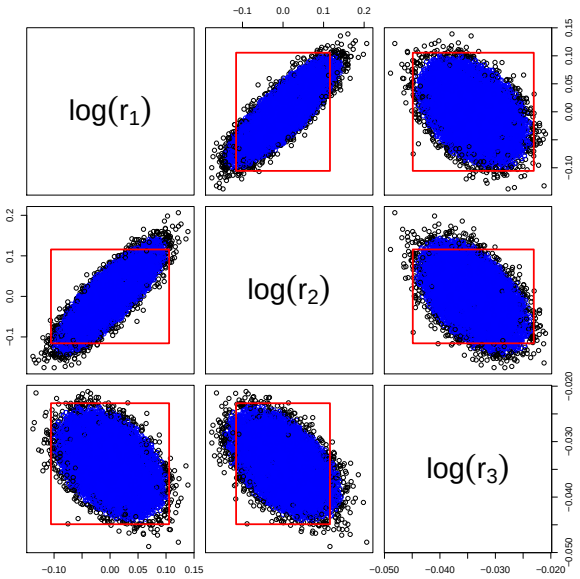


Figure 2.4: Pair plot of the joint posterior predictive distribution of $\hat{\lambda}$ when using non-informative prior distributions. The blue points are the points inside the 99%-coverage ellipsoid and the black points are the points outside the 99%-coverage ellipsoid. The red rectangles represent the limits of the hyper-rectangle test

2.4.3 Step 3. Control for consumer risks

Assume a maximum relative error of 30% is searched. 50,000 reference curves from 50,000 different plates were generated from the posterior distribution of model parameters. We then simulated one test curve from each plate, each with a different set of possible parameter ratios. The following steps were followed:

1. Draw one element of the joint posterior distribution of $\theta, \sigma_e^2, \sigma_{yMax}^2$.
2. Let $\theta_R = [yMin, yMax, C_{50}, S_{plate}]$.
3. Randomly chose r_1, r_2, r_3 and RP from the following distributions:
   - $r_1 = \exp\left(\mathcal{U}\left(\log\left(0.8\right), \log\left(1.25\right)\right)\right)$
   - $r_2 = \exp\left(\mathcal{U}\left(\log\left(0.8\right), \log\left(1.25\right)\right)\right)$

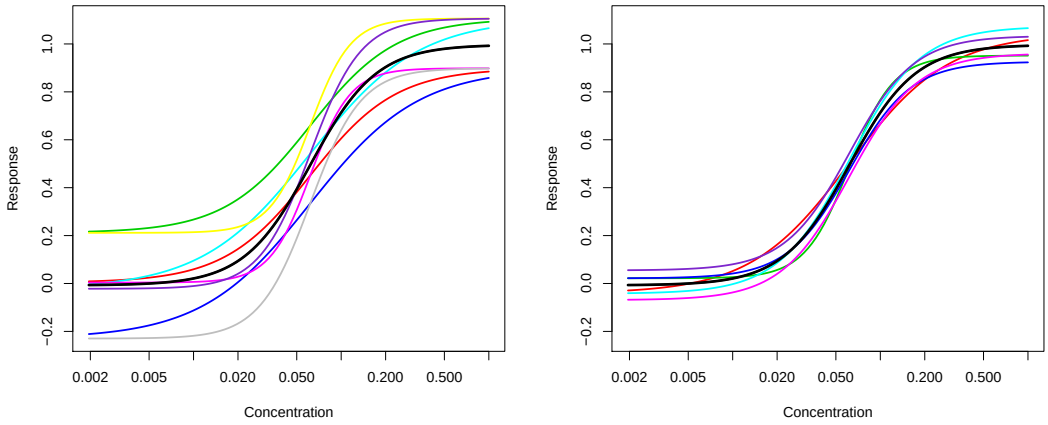


Figure 2.5: Worst acceptable curves for the hyper-rectangle (left) and ellipsoid (right) tests when using non-informative prior distributions. The black curve represents the 4PL model with the means of each marginal posterior distribution as curve parameters. The colored curves represent the sample curves for which the ratios would be equal to each of the eight corners of the hyper-rectangle (left) or each of the six summit of the ellipsoid (right).

- $r_3 = \exp\left(\mathcal{U}\left(\log\left(0.667\right), \log\left(1.5\right)\right)\right)$
- $RP = \exp\left(\mathcal{U}\left(\log\left(0.5\right), \log\left(2\right)\right)\right)$

With $\mathcal{U}\left(a, b\right)$ a Uniform distribution between a and b. These distributions were chosen to ensure that the full range of accepted λs would be contained in the simulations.

4. Let $\theta_T = [yMin_T, yMax_{plate_T}, C_{50_T}, S_T]$, with
   - $yMin_T = r_1 yMax_{plate} - r_2\left(yMax_{plate} - yMin\right)$
   - $yMax_T = r_1 yMax_{plate}$
   - $C_{50_T} = \frac{C_{50}}{RP}$
   - $S_T = S\frac{r_3}{r_2}$
5. Generate two curves using model 2.2, using θ_R for the first curve and θ_T for the second curve, with the same concentration points as the raw data and 3 replicates per concentration point for each of the curves.
6. fit Equation 2.2 using ordinary least square and calculate $\hat{\lambda}$.
7. Assess if Equations 2.3 and 2.4 are verified.
8. Compute Equation 2.5.

We report the posterior predictive distribution of the % relative error (ψ) for curves that pass parallelism only, separately for each test (Figure 2.6). The 95^{th} percentile of ψ for curves accepted by each of the two tests is presented in Table 2.2. The ellipsoid test is the only test for which $\beta^* \leq 30\%$. Using the hyper-rectangle can lead to almost 50% relative error in relative potency estimation, which is over double β^* for the ellipsoid.

Table 2.2: 95^{th} percentile of relative error in accepted curves by each test when using non-informative prior distributions.

Test	$\boldsymbol{\beta^*}$
HR	49.50%
Ellipsoid	22.62%

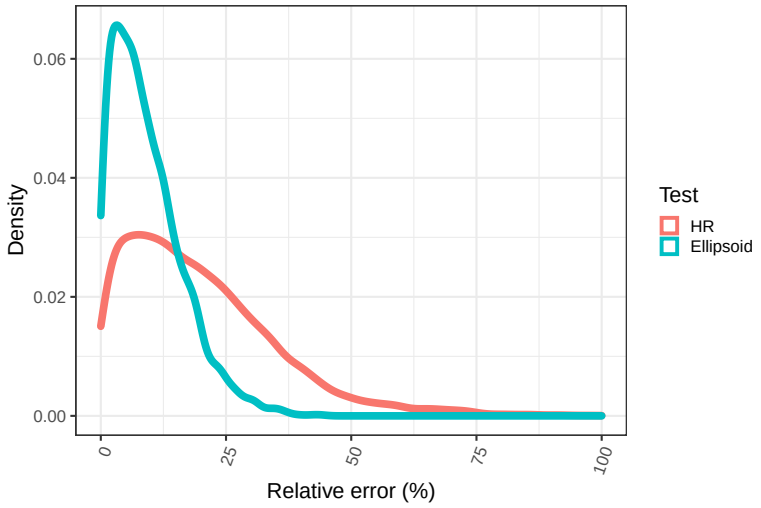


Figure 2.6: Density of ψ for curves accepted by the hyper-rectangle (HR) test and the ellipsoid test when using non-informative prior distributions.

2.4.4 Results with informative prior distributions

For the case study illustrated above, we considered that no prior knowledge was available. Often, however, a certain amount of development data is available and informative distributions can be derived. For this Section, we generated the results for the same case study, but using the following prior distributions for the model parameters:

- $yMin \sim \mathcal{N}(0.1, 0.015^2)$
- $yMax \sim \mathcal{N}(1, 0.03^2)$
- $\log(C_{50}) \sim \mathcal{N}(-2.77, 0.3^2)$
- $S \sim \mathcal{N}(2, 0.1^2)$
- $\sigma^2_{yMax} \sim Inv - Gamma(2, 0.006)$
- $\sigma^2_e \sim Inv - Gamma(30, 0.038)$

These prior distributions have been chosen arbitrarily for the purpose of this work. In practice, informative prior should always be justified using historical data and/or scientific evidence. See Klauenberg et al. (2015) for guidance on informative prior distributions in serial dilution assays [105]. The Gamma distribution is preferred as a prior for precision by some statistical software because of its role in conjugate priors for normal likelihood function [102]. We however used Stan, which uses standard deviations, making it more intuitive to use inverse-Gamma for variances.

The posterior distribution of curve parameters Φ was sampled by Hybrid MCMC using Stan, with the same model specifications as in Section 2.4.1. The trace plots of the MCMC chains are presented in Figure 2.7. As expected, the marginal posterior distributions appear less variable than they do in Figure 2.3. It is of critical importance to obtain precise estimate as all uncertainties are integrated out in the posterior predictive distribution of the test metrics.

The zone of declared similarity, both for the hyper-rectangle and the ellipsoid, was derived from Φ following the same approach as in Section 2.4.2. The estimated limits for the hyper-rectangle test are presented in Table 2.3. Exactly 99% of the calculated $\hat{\lambda}$s respected Equation 2.3.

Table 2.3: Calculated limits for the hyper-rectangle test in natural log and original scale when using informative prior distributions

Ratio	**Limits in log scale**	**Limits in original scale**
r_1	[-0.077; 0.077]	[0.926; 1.080]
r_2	[-0.107; 0.107]	[0.898; 1.113]
r_3	[-0.279; 0.279]	[0.757; 1.322]

The variance covariance matrix $\hat{\Sigma}$ of the posterior predictive distribution of λ was also calculated:

$$\hat{\Sigma} = \begin{bmatrix} 0.0007 & 0.0082 & -0.0010 \\ 0.0082 & 0.0014 & -0.0017 \\ -0.0010 & -0.0017 & 0.0084 \end{bmatrix}$$

We observed that 98.8% of the calculated $\hat{\lambda}$s respected Equation 2.4.

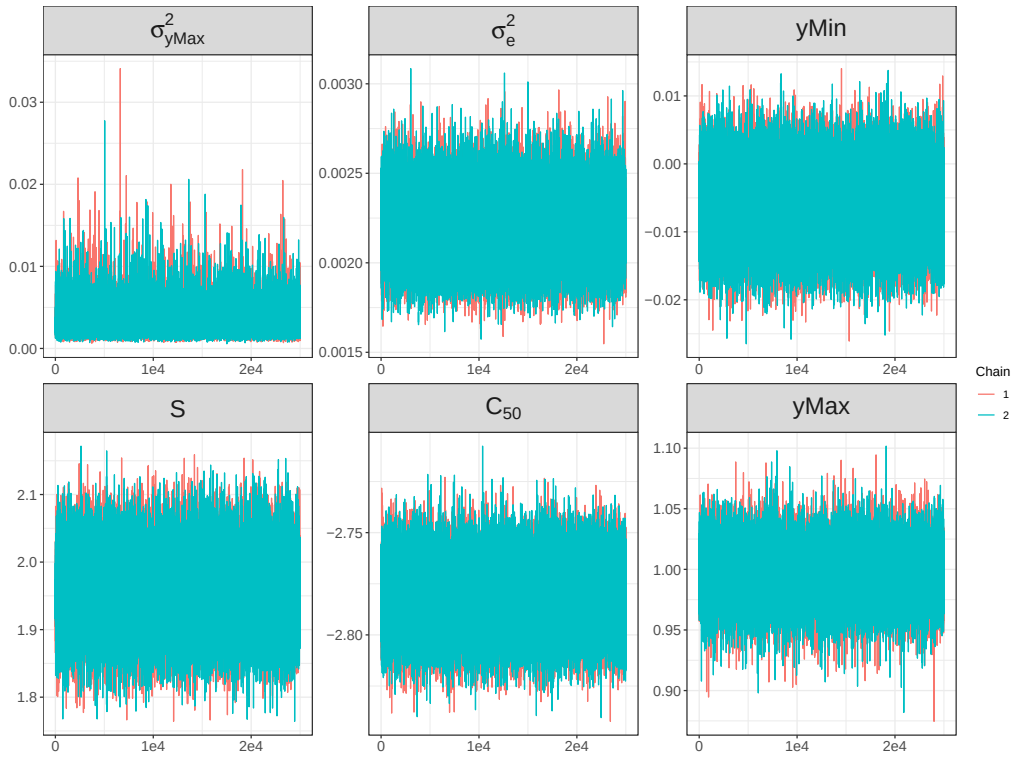


Figure 2.7: Trace plot of each model parameter when using informative prior distributions. From left to right: σ^2_{yMax}, σ^2_e, and $yMin$ on top; and S, C_{50}, and $yMax$ in the bottom

We followed the same steps as in Section 2.4.3 to estimate the consumer risk. The 95^{th} percentile of ψ for curves accepted by each of the two tests is presented in Table 2.4. The consumer risk of the hyper-rectangle test seems to be more affected by the informative prior but remains higher than the consumer risk of the ellipsoid test.

Table 2.4: 95^{th} percentile of relative error in accepted curves by each test when using non-informative prior distributions.

Test	$\boldsymbol{\beta^*}$
HR	38.20%
Ellipsoid	20.44%

2.5 Discussion

This paper presents a multivariate similarity test as well as risk-based acceptance limits derivation similarity for potency assays. The proposed test reduces the risk of high error in relative potency estimation, while also controlling for the risk of rejecting good curves. The suggested ellipsoid method is a multivariate sample suitability test rather than the common intersection-union test proposed by Jonkman and Sidik (2009) [83], and then Yang et al (2012) [84].

Hypothesis testing is used when a sample of observations is assumed to be representative of an entire population. In many cases, however, the primary interest is in the relative potency estimate, which is affected by similarity of the samples. Similarity is then a necessary condition for each sample, to ensure that the relative potency can be accurately estimated. The accuracy will depend on the estimated ratio of parameters, not the unknown true value, and so the confidence interval in that case is irrelevant. Equivalence tests based on confidence intervals are therefore not necessary for sample suitability. Instead, a zone of declared similarity, such as a hyper-rectangle or an ellipsoid containing a large portion of the posterior predictive distribution of the test statistics in case of 'true' similarity, can be derived and estimates can be compared to said zone of declared similarity. Berger and Hsu (1996) advised against multivariate testing for bioequivalence and prefer intersection-union tests [86]. Their main reasoning seems to be that $\hat{\Sigma}$ is difficult to derive, and the similarity zone is therefore poorly established. In the case of similarity testing, we agree that there is currently no exact way to derive it analytically. However, we showed a way to compute the joint posterior predictive distribution of the parameter's ratio estimates using Bayesian methodologies, making the multivariate zone of declared similarity easily derivable. In the context of Berger and Hsu's paper, an additional reason-

ing to avoind multivariate is that to demonstrate bio-equivalence, confidence intervals are required. Our approach is focused on point estimate testing for sample or system suitability in potency assays, and may not be directly generalizable to any multivariate test.

We ignored heteroscedasticity for the sake of simplicity. A common way to address heteroscedasticity is by fitting the 4PL model using weighted least square [106]. This is usually a poor choice in bioassay models, and repeated measure models in general; nonlinear mixed model are usually preferred [80, 81]. Our proposed methodology works similarly with or without extra variance components.

Here, the laboratory risk is calculated for curves that have exactly equal θs. Some may argue that this is too conservative and that slight departure in θs is acceptable. Following our approach will lead to α increasing as true θs get more different. For a lab desiring a fixed α when $\theta_T \approx \theta_R$, a simple solution is to include slight differences in θs when generating reference curves in Step 1 and calculate $p\left(r|y, \mathcal{I}, \sigma^2, \theta_R, \theta_T \approx \theta_R\right)$. Step 3 in that case is particularly important as this could increase the error in relative potency estimation.

If the consumer risk (i.e. the error in relative potency estimation) is deemed too high, different corrective actions can be undertaken. Increasing the number of replicates is an expensive but easy way to lower the uncertainty of the test statistics, reducing the volume of the three-dimensional space under true parallelism, and increasing the precision of the relative potency estimation. Another option is to improve the assay design, either by expanding the range of concentrations or increasing the number of dilutions within the current range. If those actions, that can easily be evaluated through simulations, are not possible or not sufficient to reduce relative potency error, final options are accepting a higher lab risk to tighten the limits or improving the assay itself. Process optimization methodologies can easily be applied to assay development to find the optimal assay parameters (temperature, incubation time, plate washing time, etc.) [107]. Note that applying those methodologies from the get-go can prevent having to spend time and money to improve an assay later, when it is often costlier and challenging to do so [93, 108].

A limitation of the ellipsoid approach is that Bayesian Markov Chain simulations are required to derive the ellipsoid for every new assay, which may be

not obvious computations. However, once $\hat{\Sigma}$ is calculated, the routine similarity test can easily be implemented in bioassay analysis software or even Microsoft Excel and doesn't rely on simulations anymore. If no prior knowledge about the curve parameters is available, bootstrap or jackknife can be used instead of MCMC simulation to simplify the approach. The bootstrap distribution can be viewed as an approximate posterior distribution [109]. However, as shown in Section 2.4.4, informative prior distributions may help narrowing the zone of declared similarity and therefore lower the consumer risk. This may be viewed an indication that updating the acceptance limits periodically is recommended. When doing so, the information gained before and during the validation can be used to construct informative prior distributions. Whether MCMC, bootstrap, or other resampling method is used to estimate the distribution of the log(ratios) of future pairs of reference curve, the accuracy of the estimates relies on the assumption that the observed data (and, if used, the informative prior distributions) are a good representation of future results. Acceptance limits should be re-evaluated in case of any change in the assay design or performances.

Another limitation is that, to derive an ellipsoid, the multivariate (log-)normality assumption is necessary. Rather than relying on multivariate normality tests such as Mardia test for multivariate normality [110], we recommend careful visual inspection of the posterior predictive distribution of the ratio estimates to determine if the log-normality assumption is reasonable. If the multivariate normality assumption is not reasonable, other elliptical distributions and their coverage ellipsoid must be investigated [100,111,112].

An important note is that, while testing ratios simultaneously appears to be a good approach than testing them separately to assess similarity, we still recommend reporting and monitoring each ratio separately. If investigation is needed related to non-similarity or loss of potency, trending the ratios individually can provide useful information. Additionally, while MCMC simulations are a good alternative to actually generating many plates filled with reference products, performing some real reference-to-reference comparisons is recommended to ensure that the MCMC simulations are a good representation of reality.

Finally, this manuscript aims to present the multivariate approach as a reasonable alternative to the marginal tests. However, a smaller consumer risk for a single case study is not enough to demonstrate the general superi-

ority of the proposed approach over the hyper-rectangle test, or other tests proposed in the literature. Extensive comparisons will be the object of future research.

Chapter 3

Effect of a Statistical Outlier in Potency Assays

This Chapter is based on the article titled "Effect of a statistical outlier in potency bioassays", published in *Pharmaceutical Statistics* in November 2018 [113]. The vocabulary has changed to be consistent with the rest of the dissertation. The similarity test used has also changed: the paper uses a hyper-rectangle test for parallelism, while the chapter uses the ellipsoid test presented in Chapter 2.

3.1 Introduction

For most biotherapeutic products and vaccines, the potency must be measured and compared with specification limits prior to batch release. Potency is defined in ICH Q6B as the measure of the biological activity of a sample using a suitably quantitative biological assay (bioassay), and biological activity is defined as the specific ability or capacity of the product to achieve a defined biological effect [114, 115].

Potency assays may be used to measure the biological activity or binding affinity of a test sample relative to that of a reference preparation [47, 115]. For example, in the manufacture of a biotherapeutic or vaccine, the potency of material from a new batch may be calibrated relative to the potency of reference batch material. The relative potency (RP) is calculated from

the concentration-response functions of two products. Two products exhibit constant RP whenever the test product acts as a dilution or concentrate of the reference product. Graphically, two concentration-response curves that exhibit constant RP differ only by a shift on the log-concentration axis. A calibration based on a constant RP value may only be performed when concentration-response functions of the two products are declared similar (or parallel) through statistical testing [58, 78].

In practice, outliers are caused by unexpected sources of variation, such as from pipetting error, dilution error [99], wrong incubation time, or other mistakes in the lab procedures. When statistical outliers are present in the data and no action is taken to remove them, the test of similarity may be compromised, and the RP value may be estimated with bias. A statistical outlier is a value that appears to be inconsistent with the other observations in the available data [116]. An outlier in a potency assay may result from several causes, such as experimental error, variability in experiment material, or variability in experiment technique [117]. The United States Pharmacopeia (USP) <1032> recommends that bioassay data should be screened for outliers before RP analysis. Schofield writes "Some laboratories will exclude the individual outliers while others may exclude a dilution or dilutions associated with the outliers. The lab should assess the impact of exclusion of the outlier on similarity testing and relative potency determination. [118]" The actual consequence of an outlier in a potency assay, however, has never been quantified. The aim of this Chapter is to assess, using simulation studies, the effects of various types of outliers on both RP estimation and similarity testing. In the presence and absence of outliers, the Monte Carlo probability of false rejection of similarity when the reference and test curves are exactly similar is closely examined. For curves that are declared similar, the effect of outliers on RP estimation is investigated. In addition, the Monte Carlo probability of false acceptance are examined for a case in which the reference and test curves are not similar.

In Section 3.2, three types of outliers are given and illustrated graphically. In Section 3.3, a method of similarity testing and a method for estimating RP are shown for a set of test and reference concentration-response data sets. The simulation setup is show in Section 3.4 with results given in Section 3.5. Finally, Section 3.6 contains a discussion and advice about outlier removal before similarity testing and calculation of RP.

3.2 Outlier Types

Most scientific publications regarding outliers refer only to a single observation outlier [119, 120]. In the case of potency assays, however, several types of outliers are possible. In this work, we consider an experimental design in which the reference and test concentration-response curves are collocated on each of three 96-well plates. The three sets of curves may be mutually independent, but the assay measurements within each of the three plates may be correlated. Different outlier scenarios are considered in this Chapter.

1. Single observation outlier: a single measurement is excessively distant from the other values (Figure 3.1, left).

2. Concentration point outlier: all the replicates at one concentration level exhibit a different behavior than the values at the other concentration points (Figure 3.1, right)

3. Whole curve outlier: one of the curve replicates is affected by a manipulation error. It can affect the dilution factor (Figure 3.2, left), one of the asymptotes (Figure 3.2, middle), or both asymptotes, inducing a vertical shift (Figure 3.2, right).

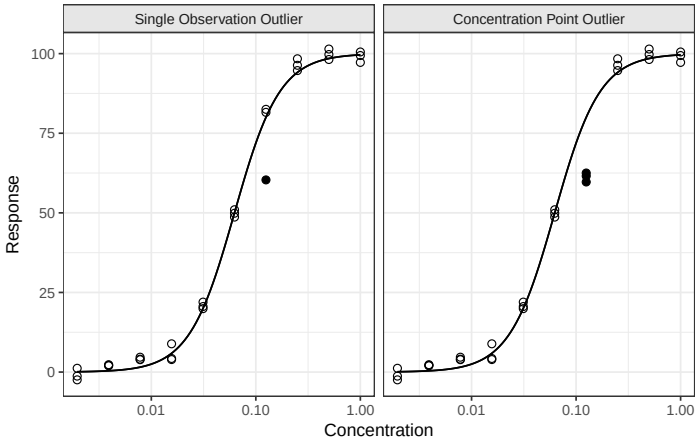


Figure 3.1: Examples of single observation outliers (left) and concentration points outliers (right)

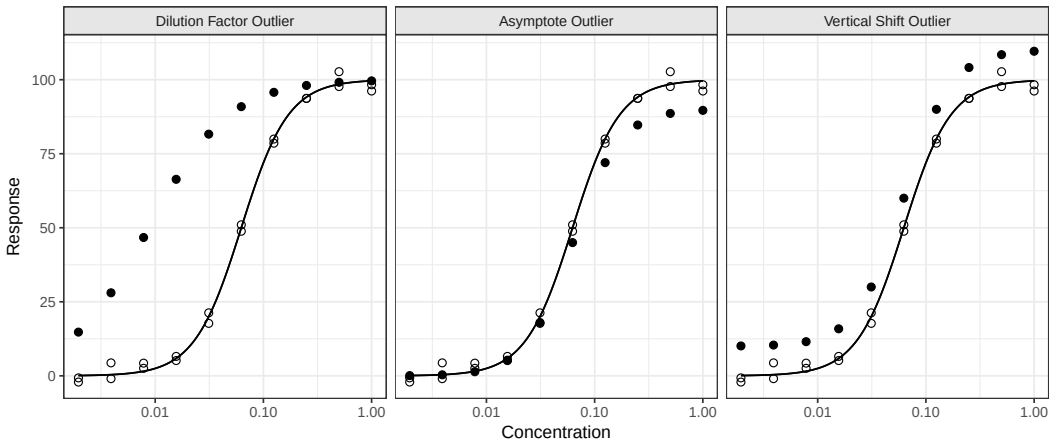


Figure 3.2: Examples of whole curve outliers with an effect on the dilution factor (left), the upper asymptote (middle) or both asymptotes (right)

3.3 Similarity Test and Calculation of RP

A commonly used model for potency assay concentration-response data is the four-parameter logistic (4PL) curve. The usual parametrization was proposed by Rodbard and Hutt [50], described in Section 1.2.1:

$$y_{ij} = yMax_i + \frac{yMin_i - yMax_i}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^{S_i}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{3.1}$$

For the sake of simplicity, we only consider the simple case where the errors ϵ_{ij} are independent and identically distributed. It is not uncommon in potency assays to observe errors that are not independent or normally and identically distributed. Such data require more sophisticated modeling of the variance structures.

Both the USP [47] and the European Pharmacopeia (EP) [53] require assessment of similarity before the computation of relative potency. The recent literature has promoted equivalence testing for similarity as an alternative to difference testing [82] and is also recommended by the USP <1032> [47]. In the case of 4PL curves, several authors propose to compare the lower asymptotes, upper asymptotes, and growth rates between both reference and test curves via equivalence testing [83,84,99]. In this Chapter, we use the ellipsoid test proposed in Chapter 2.

The RP is computed for data that are declared similar. To perform the RP calculation, both reference and test curve data are jointly modeled using the following formula:

$$y_{ij} = yMax + \frac{yMin - yMax}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^{S}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{3.2}$$

Equation 3.2 is an adaptation of Equation 3.1 where $yMin$, $yMax$ and S are the same for both reference and test curves. For parallel curves, the true RP is the ratio of unknown C_{50} model parameters. The estimated RP is the ratio of estimated C_{50_i}s.

3.4 Simulation Setup

For each simulated run, three replicates each of a reference and a test product concentration-response curve were simulated via Equation 3.1, across ten dilution steps. The ten concentration points are given as $[1, 0.5, 0.25, \ldots, 0.002]$ with a ½dilution factor. Highly precise assays ($\sigma = 2$) and highly variable assays ($\sigma = 15$) were examined. The parameters for the reference product curve were held fixed at $(yMin = 0, yMax = 100, c_{50}, S = 2)$. The reference curve was always generated in an outlier-free manner. The parameters for the test curves vary in order to analyze the results both in the case of true parallelism and true non-parallelism. We assume that no observation falls above or below the natural bounds of the assay signal (i.e., saturation point and limit of detection). To demonstrate the effect of outliers on parallel curves, the test product curves were generated with parameters $yMin$, $yMax$, and S set exactly equal to those of the reference product curve. The test product C_{50} was varied as 50%, 100%, and 200% of the reference product C_{50}, leading to a true RP of 200%, 100% and 50%, respectively. There are an infinite number of ways to generate non-parallel curve data. To avoid overwhelming the reader, the effect of outliers on non-parallel curves is demonstrated by purposely generating the test curve with an asymptotic outlier (see middle panel of Figure 3.2). For non-parallel curves, the test product data was generated with parameters equal to those of the reference product curve, except $yMax = 100 - 3\sigma$. The results from the non-parallel simulations are illustrative and should not be considered comprehensive. No recovery results are presented for these scenarios because the true RP does not exist in case of non-parallel curves.

Outliers were included in the test curve as described in Section 3.5 for each type of outlier defined in Section 3.2. Each scenario was evaluated 10,000 times.

Because concentrations were chosen to be symmetrical to the C_{50}, the effect of an outlier on the model is also symmetrical around the C_{50} of the reference curve and so only positive outliers are considered in this study. That is, in our scenarios, a positive outlier at the $(5-d)^{th}$ dilution step, has the same effect as a negative outlier at the $(5+d)^{th}$ dilution step.

After model fitting was performed via nonlinear least squares, the effect of outliers on similarity testing was assessed as the proportion of curves for

which the similarity is not declared (see Section 3.5.1). If the data meet with the similarity test requirements, the RP was estimated. We refer to the ratio of the estimated RP to the true RP as recovery. Recovery is expected to be 100%. Because simulated results for the parallel-curve scenarios only weakly relied on the test-curve C_{50} value, only the 100% expected RP case is shown. The results for 50% and 200% RP are presented in Appendix B. The non-parallel curve scenarios are not similar and so there is no expected percent recovery value. The estimates of RP appear to be log-normally distributed, and so the geometric mean of the recovery across the simulations is reported. Note that, because only positive outliers were generated, their effect increases the percent recovery. Data with only negative outliers would have the equivalent opposite effect to decrease percent recovery.

3.5 Simulation Results

3.5.1 Similarity Test Acceptance Criteria

Let:

$$r_1 = \frac{yMax_T}{yMax_R}$$
$$r_2 = \frac{yMax_T - yMin_T}{yMax_R - yMin_R}$$
$$r_3 = \frac{(yMax_T - yMin_T)S_T}{(yMax_R - yMin_R)S_R}$$

For each generated pair of curves, r_1, r_2 and r_3 were calculated. The indifference zone for the ellipsoid similarity test (see Chapter 2 for details) is the ellipsoid containing 95% of $\hat{\lambda} = [\log(\hat{r_1}), \log(\hat{r_2}, \log(\hat{r_3}]^t$ of the 10,000 simulated curves under true parallelism with no outliers with RP=1, separately for each σ.

For a pair of curves to be declared parallel, it has to verify:

$$\hat{\lambda}^t \hat{\Sigma} \hat{\lambda} \leq q_{0.95,3} \tag{3.3}$$

where $\hat{\Sigma}$ is the variance covariance matrix and $q_{0.95,3}$ is the 95^{th} percentile of $\hat{\lambda}$ from the 10,000 curves that were used to derive the indifference zone.

3.5.2 Single Observation Outlier

To generate a single outlier, the factor $k\sigma$ (a multiple of the measurement variability) is added to a single observation with $k = 4, 6$, and 10. Because outliers can occur at any dilution step, the impact was evaluated at each dilution step in the simulations.

Figure 3.3 shows the probability of rejecting similarity when the curves are actually parallel (left panel) and non-parallel (right panel), as a function of the outlier severity for the case of a single observation outlier. For exactly parallel curves, when $k = 0$ (no outliers), the similarity test rejection rate for parallel curves is 5%, per construction. When the assay variability was low ($\sigma = 2$), the impact of outliers on similarity testing depended on its severity and position on the curve, depending on whether concentrations to the left and right provided support for the particular portion of the curve. For example, an outlier at concentration=1.0 had a larger impact on similarity testing than an outlier at concentration = 0.5 because there is a reduced support for the curve shape at the largest concentration. As expected, non-parallel curves are rejected almost 100% of the time when $k = 0$ (no outliers). Because non-parallelism was constructed with $yMax_T < yMax_R$, outliers at the low concentrations have little effect on the rejection rate, but positive outliers at the high-concentration in the test curve data lead to an overestimation the upper asymptote, which "fixes" the problem and can lead to false acceptance of non-parallel curves. Note again that different scenarios of non-parallelism will lead to different results. When the assay variability was high ($\sigma = 15$), the effect of a single observation outlier in the lower asymptote or at the highest concentration was similar as when the assay variability was low ($\sigma = 2$), but not exactly the same concentrations where most negatively affected.

Figure 3.4 shows the geometric mean of the recovery for cases that were declared similar in the presence of a single observation outlier at a given dilution step. An outlier in the asymptote (highest and smallest concentrations) has virtually no impact on the RP estimation, while an outlier in the middle of the curve can have a major impact, depending on the residual standard

deviation and outlier severity. For precise assays ($\sigma = 2$), even the most extreme outliers did not induce a relative error in recovery of parallel curves above 10%, while outliers in imprecise assays had a stronger impact, up to 60% relative error.

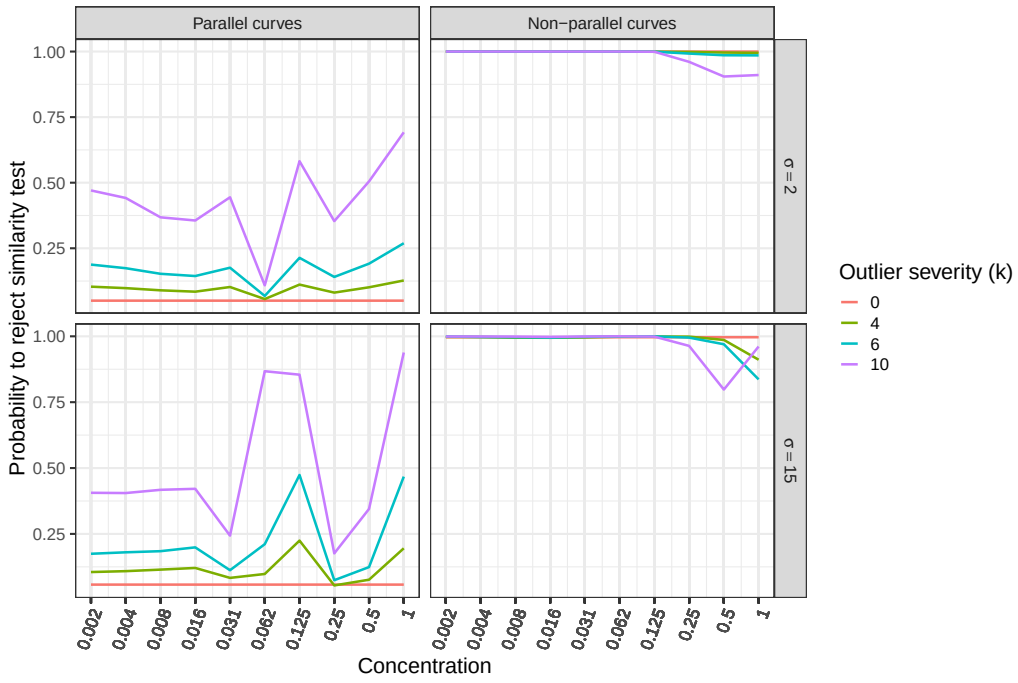


Figure 3.3: Single observation outlier: effect on similarity testing when true RP = 100%

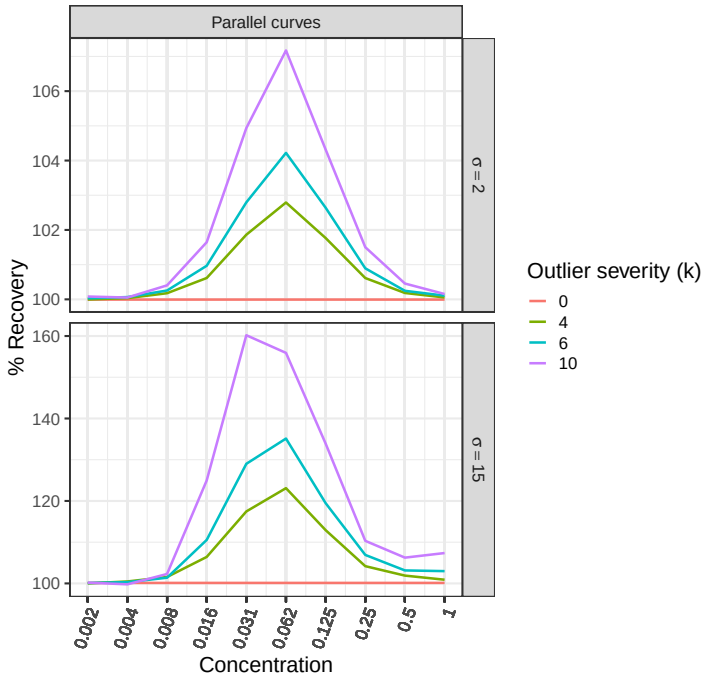


Figure 3.4: Single observation outlier: geometric mean of the recovery when true RP = 100%

3.5.3 Concentration Point Outlier

For concentration point outliers, the outlier generation process is the same as for single observation outliers, except that the factor $k\sigma$ is added to all the replicates at the selected dilution step. To avoid a wholesale increase in the parallelism rejection rate, less severe outliers were generated ($k = 2, 3,$ or 4) compared to those of Section 3.5.2.

Figure 3.5 shows the probability of rejecting similarity with a single concentration outlier follows a pattern similar to that of the single point outlier. Because all three observations from a single concentration are pushed away from the true curve, less severe outliers can have a significant impact on rejection of similarity. Figure 3.6 shows that the effect of concentration point outliers on RP estimation is also similar to that of single point outliers, but more severe.

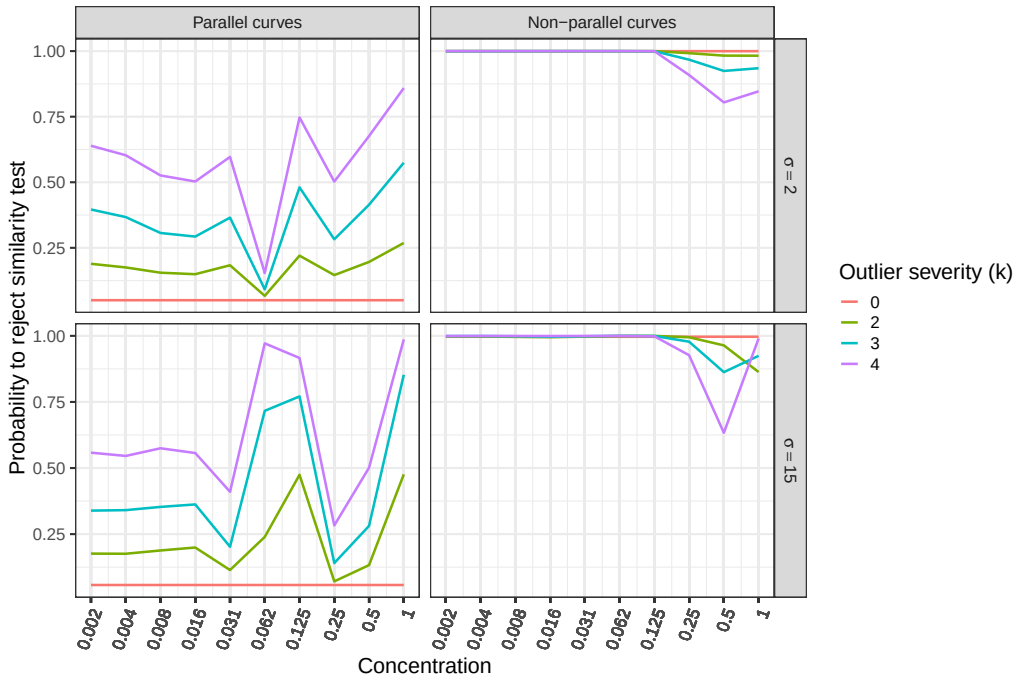


Figure 3.5: Concentration point outlier: effect on similarity testing when true RP = 100%

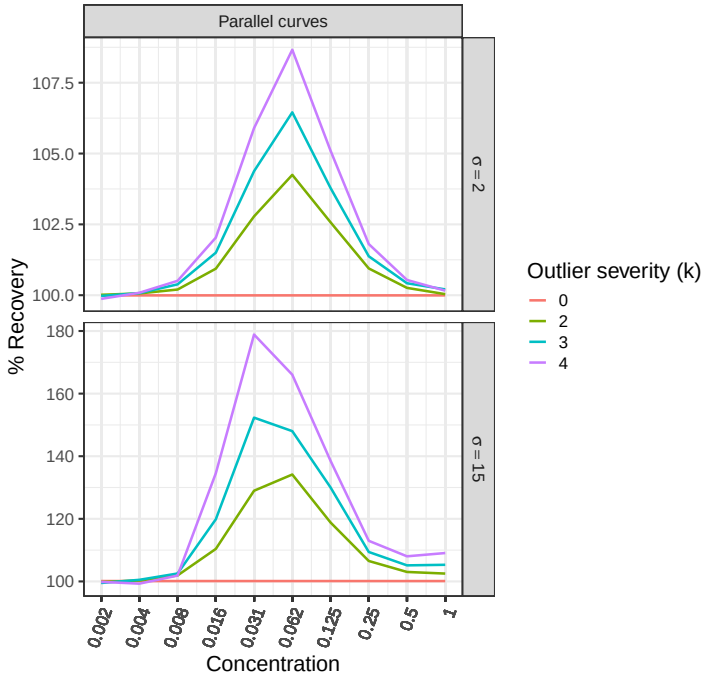


Figure 3.6: Concentration point outlier: geometric mean of the recovery when true RP = 100%

3.5.4 Whole Curve Outlier

In this section, three types of whole curve outliers were generated, namely:

(a) Replacing the serial dilution factor of 2 with 2.5 for one of the replicates (dilution factor outlier). As a consequence, the true concentration is different than the observed concentration for one of three whole-curve replicates.

(b) Changing $yMax = yMax + 10$ for one of the replicates (upper asymptote outlier).

(c) Changing $yMax = yMax + 10$ and $yMin = yMin + 10$ for one of the replicates (vertical shift outlier).

Because these generated outliers do not depend on the measurement variability, it is expected that the effect of the outlier is more important for precise assays compared to imprecise assays.

The effect of a whole curve outlier on the similarity tests is shown in Figure 3.7. Because outlier generation is not a multiple of σ, the higher the measurement variability, the less the outlier affects similarity equivalence testing. Such outliers therefore have low to no impact on similarity testing when $\sigma = 15$. When $\sigma = 2$, a shift in one or both asymptotes in one out of three test curves has a profound effect on similarity testing for truly parallel curves. For our scenario of truly non-parallel curves, a positive upper asymptote outlier lead to mistakenly accept over 50% of the curves. Changing the dilution factor of one test curve had only a modest impact on similarity testing

Figure 3.8 shows the geometric mean of the recovery for the evaluated relative potencies and measurement variabilities in case of whole curve outliers. The geometric mean recovery of truly parallel curves does not appear to be affected by the measurement variability.

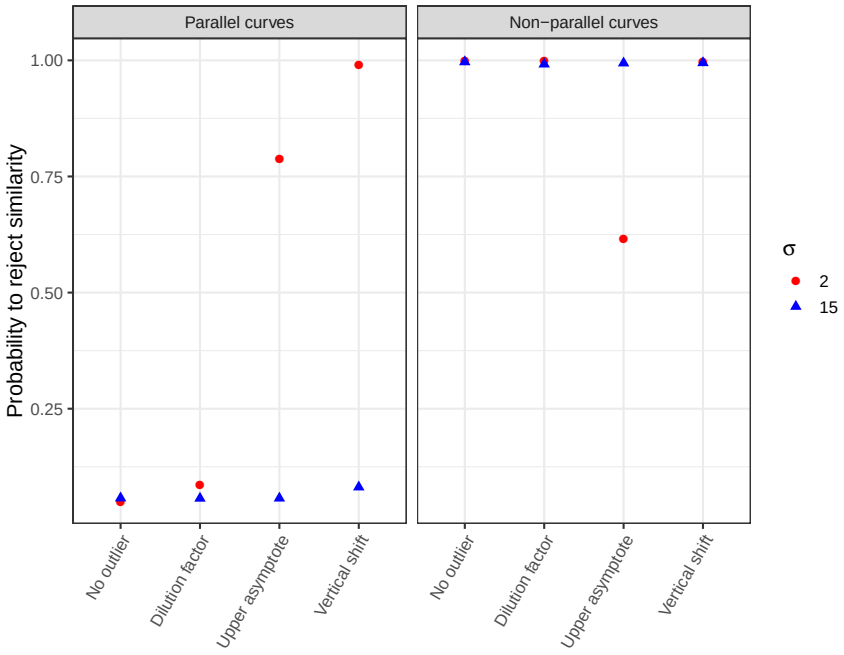


Figure 3.7: Whole curve outlier: effect on similarity testing when true RP = 100%

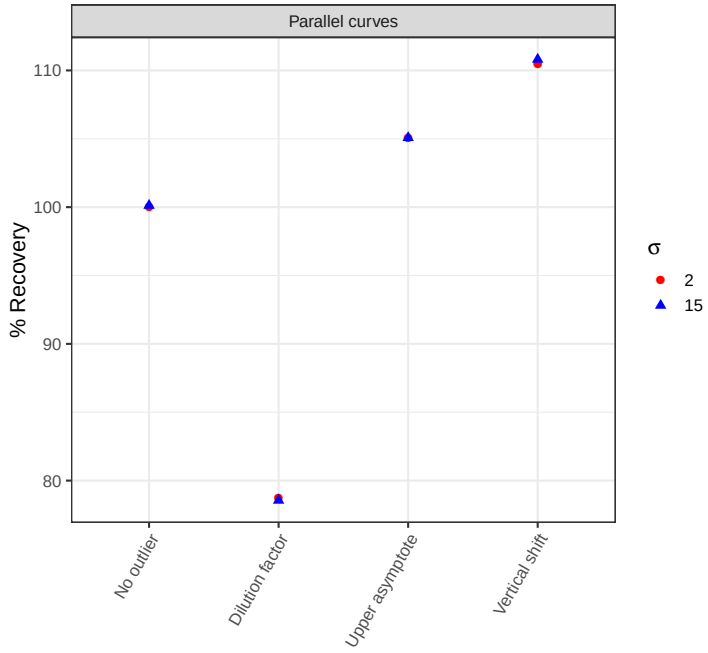


Figure 3.8: Whole curve outlier: geometric mean of the Recovery when true RP = 100%

3.6 Discussion

For potency assays, USP<1032> recommends screening for outliers. Until now, the effect of an outlier on the quality of the RP has not been quantified to confirm the utility of outlier testing. This chapter examined the effect on the performance of parallelism testing and the magnitude of the error in RP estimation induced by the presence of an outlier in the data set for various outlier types that may be expected to occur in practice, including whole curve outliers.

The simulation study shows that, for single outliers and single-concentration outliers, the effect of the undetected outlier on RP estimation is a function of the outlier magnitude relative to the assay precision and outlier position on the curve. Very precise assays concentration-response curves with a single outlier or concentration-point outlier tend to produce nearly unbiased RP values when the curves are truly parallel; however, this may come at the cost of a high similarity testing rejection rate. As the assay precision diminishes, the single outlier and single-concentration outlier may lead to both a higher similarity test rejection rate for truly parallel curves as well as a gross bias in RP estimation. Whole-curve outliers are universally worrisome, potentially leading to similarity testing failure or high RP bias. When the test product is not parallel to the reference, outliers can lead to false declarations of similarity. We thus agree with the USP recommendations to screen for outliers and to test for similarity before estimating RP.

Chapter 4 focuses on the performance of various outlier test methods, including proposals for whole-curve outlier testing. Future research will include the assessment of the effect of outliers when the residuals are not independent, identically and normally distributed.

The effect of an outlier depends on its location on the curve. Therefore, a concentration design that lacks supports in either asymptote or in the ascending portion of the curve may increase the negative effects that outliers can have on either similarity assessment or relative potency estimation. Similarly, poor design may increase the negative effect of outliers. See Chapter 5 for design considerations.

Chapter 4

Comparison of Outlier Tests for Potency Bioassays

This Chapter is based on the article titled "Comparisons of outlier tests for Potency Bioassays", published in *Pharmaceutical Statistics* in November 2019 [121]. The vocabulary has changed to be consistent with the rest of the dissertation. The similarity test used has also changed: the paper uses a hyper-rectangle test for parallelism, while the chapter uses the ellipsoid test presented in Chapter 2.

4.1 Introduction

ICH Q6B defines the potency of a biotherapeutic product or vaccine as the measure of its capacity to achieve a defined biological effect [114, 115]. Measuring potency directly is impossible. Instead, one must measure it relative to that of a reference preparation [3, 47]. In the manufacturing of biotherapeutics, the potency of a new batch is calibrated relative to the potency of a reference standard by comparing the concentration-response curves of the two samples. For a given response value y_0, the relative potency (RP) of the batch is calculated as the ratio of the concentrations from the reference standard and batch samples. If the RP is constant or nearly constant for all possible response values, then the two curves shapes are equivalent, except for a shift along the concentration axis. In such a case, the two curves are

declared to be similar (or parallel) and the batch and reference standard are said to exhibit constant relative potency. In the context of biotherapeutic manufacturing, the computed RP value is only useful for calibration if those batch and reference standard curves are declared to be similar (parallel) [78].

When statistical outliers are present in the measured concentration-response data for the batch and/or the reference standard, Sondag et al. (2018) show that similarity testing may be compromised, and RP estimates may be biased [113] (see Chapter 3). These findings agree with the USP <1032> recommendations to screen for outliers before RP analysis [47]. Some outlier testing guidance is provided in USP<1010> and, while there is no direct claim in that chapter that its outlier tests are ideal for bioassay [122], the USP<1010> recommendations represent a broad swath of common, simple outlier tests. Through computer simulation, the utility of the USP<1010> recommendations were evaluated in terms of sensitivity and specificity to detect outliers and the subsequent effect of the removal of declared outliers on similarity testing and RP estimation. In addition to the recommended methods in USP<1010>, we examined the robust outlier (ROUT) detection method of Mutulsky and Brown [120] as well as two novel methods that we propose.

As in Chapter 3, three types of outliers are examined:

1. Single observation outlier: a single measurement is excessively distant from the other values.

2. Concentration point outlier: all the replicates at one concentration level differ from the values at the other concentration points.

3. Whole curve outlier: one of the curve replicates is affected by a manipulation error. It can affect the dilution factor, one of the asymptotes, or both asymptotes, inducing a vertical shift.

Although outlier detection may be performed in conjunction with a litany of concentration-response curves, as in Chapter 3, the assumed mean model for potency bioassay concentration-response data is the four-parameter logistic (4PL) curve. The usual parametrization was proposed by Rodbard and Hutt [50], described in Section 1.2.1:

$$f(\theta, x) = yMax + \frac{yMin - yMax}{1 + \left(\frac{x}{C_{50}}\right)^S} \tag{4.1}$$

USP <1010> [122] suggests three tests for outlier detection: The Hampel's Rule [123], the Extreme Studentized Deviate test [124,125], and Dixon's Q test [126]. In this Chapter, the performance of these tests is assessed for potency bioassay concentration-response functions through the 4PL curve. In addition to the three outlier tests suggested by USP<1010>, Motulsky and Brown's outlier detection method (ROUT) [120] and a novel robust prediction interval (RPI) method based on Huber weighting are also evaluated [127]. Computer simulation studies were run to assess the performance of the five outlier detection methods. For whole-curve outliers, a novel approach is separately proposed.

Our computer simulations mimicked a typical RP-calculation scenario. Given concentration-response data from a reference standard (Ref) and a test batch (Test) samples, model (1) is fitted separately to each sample, resulting in estimates for the reference standard and test model parameters, θ_R and θ_T. Before estimating RP, parallelism testing is performed to determine if the shapes of the two concentration-response curves are similar. The test may be performed by comparing the reference and test parameters $yMin$, $yMax$, and S, as described in Chapter 2. When no outliers are present in the data, the lab risk to reject truly parallel curves is gauged to 5%. If the two curves are declared similar, the following model is jointly fitted to the reference and test curve data:

$$f(\theta, x) = yMax + \frac{yMin - yMax}{1 + \left(\frac{x}{C_{50_i}}\right)^S} \tag{4.2}$$

In Equation 4.2, the parameters $yMin$, $yMax$, and S are common to both reference and test curves, but the C_{50} parameter is estimated separately for $i = R, T$. The RP is calculated by dividing C_{50_R} by C_{50_T}.

4.2 Outlier Tests

Outlier testing is performed on each curve (R, T) individually. Data from a single curve is assumed as $y_j = f(\theta, x_j) + \epsilon_j$, where $f(\theta, x)$ is defined in equation 4.1 and ϵ_j $\mathcal{N}(0, \sigma^2)$, $j = 1, 2, \ldots, n$. The j^{th} residual is given as $r_j = y_j - f\left(\hat{\theta}, x_j\right)$ where $\hat{\theta}$ is an estimate of θ. Let $\check{r} = median(\{|r_j|\})$. We define the median absolute deviation (MAD) of the residuals by $\check{r} \times 1.4826$. For normally-distributed errors, the MAD is an estimate of the standard deviation that is robust to outliers. All outlier testing is performed at a 1% significance level.

4.2.1 Tests for Single Observation and Concentration Point Outliers

In Chapter 3, we define a single observation outlier as a single measurement excessively distant from the other values (Figure 1, left panel). A concentration point outlier refers to the set of replicate measurements from the same concentration level at which all the replicates exhibit a different behavior (in the same direction) from the values at the other concentration points (Figure 4.1, right panel). Four outlier tests from the literature and one novel outlier test are detailed in this section.

USP<1010> recommends the use of three relatively simple outlier tests: Hampel's rule, the generalized extreme studentized deviate (ESD) test, and Dixon's Q test. For these three tests, it is assumed that θ is estimated by $\hat{\theta}$ via ordinary least squares.

Hampel's rule identifies r_j as an outlier if $\frac{|r_j - \check{r}|}{MAD} > 3.5$. Essentially, Hampel's rule declares as outliers any observations that are more than 3.5 standard deviations (MADs) from the center of the distribution (median). While Hampel's rule benefits from simplicity, the critical value of 3.5 is not adjusted for sample size and so the test size is not controlled [123].

For the generalized ESD test, let $\rho = \frac{\max |r_j - \bar{r}|}{\hat{\sigma}}$, where $\bar{r}$ and $\hat{\sigma}$ are the sample mean and standard deviation of the residuals. The value r_j that maximizes $|r_j - \bar{r}|$ is declared an outlier if ρ exceeds the critical value given

by

$$\upsilon_1 = \frac{(n-1)t_{p,n-2}}{\sqrt{\left(n-2+t^2_{p,n-2}\right)n}}$$

where $t_{p,\nu}$ is the $100p^{th}$ percentile of a Student's T distribution with ν degrees of freedom and $p = 1 - \frac{\alpha}{2n}$ with $\alpha = 0.01$, the pre-defined significance rate. The value r_j that maximizes $|r_j - \bar{r}|$ is removed from the set of residuals and the entire process is repeated from the remaining $n-1$ residuals. At the k^{th} step (at which point, $k-1$ outliers have been declared), the test statistic ρ is evaluated against the critical value

$$\upsilon_k = \frac{(n-k)t_{p,n-k-1}}{\sqrt{\left(n-k-1+t^2_{p,n-k-1}\right)(n-k+1)}}$$

with $p = 1 - \frac{\alpha}{2(n-k+1)}$. The process halts when $\rho \leq \upsilon_i$. While more complex than Hampel's rule, the ESD test does adjust for the sample size and so test size is controlled by α [125].

The last recommended method in USP<1010> is a generalized Dixon's Q test. Denote the j^{th} rank-order value of the residuals by $r_{[j]}$ so that $r_{[1]} \leq r_{[2]} \leq \cdots \leq r_{[n]}$. The test statistic for Dixon's depends on the sample

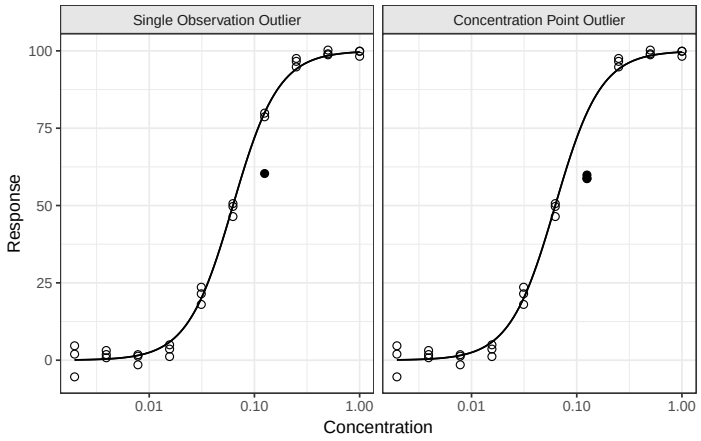


Figure 4.1: Example of single observation outliers (left) and concentration points outliers (right)

size [126]. To test the minimum (most negative) residual as an outlier for $n = 30$, let $Z_{[1]} = \frac{r_{[3]} - r_{[1]}}{r_{[n-2]} - r_{[1]}}$. To test the maximum (most positive) residual as an outlier, let $Z_{[n]} = \frac{r_{[n]} - r_{[n-2]}}{r_{[n]} - r_{[3]}}$. The critical value for both $Z_{[1]}$ and $Z_{[n]}$ is calculated by taking the 99% quantile of $\frac{\omega_{[n]} - \omega_{[n-2]}}{\omega_{[n]} - \omega_{[3]}}$, where the ω_j are independent standard normal random variables. For example, at the 1% significance level, with a sample size of N=30, the critical region for both the minimum and maximum residuals $Z_{[1]}$ and $Z_{[n]}$ is > 0.457. To use the standard Dixon's Q test, we are naively assuming that the r_j (which are correlated) are independent, identically-distributed univariate normal random variables. Dixon's test is initially applied to the full data set (n observations). If an outlier is detected, the observation is declared an outlier and removed. Dixon's test is repeated on the remaining observations $n-1$. This procedure is applied with new calculated critical values at each step until no additional outliers are detected. The generalized Dixon's Q test procedure approximately controls the test size.

Motulsky and Brown derived a robust outlier (ROUT) detection method that fits Equation 4.1 to the data assuming Cauchy-distributed errors in place of Gaussian-distributed errors [120]. A Cauchy likelihood is fitted to the data by maximizing

$$\prod_{j=1}^{n} \frac{1}{\pi\gamma\left[1 + \left\{\frac{y_j - f(\theta, x_j)}{\gamma}\right\}^2\right]}$$

with respect to (θ, γ), where γ is a scaling parameter. The robust standard deviation of the residuals ($RSDR$) is defined as the 68^{th} percentile of the absolute value of the Cauchy-fitted residuals, multiplied by $n/(n-4)$. Standardized residuals are $s_j = r_j/RSDR$. An observation is declared an outlier if the absolute standardized residual value is large by examining the p-value$= 2P(T > |s_j|)$, where T follows a central t-distribution with $n-4$ degrees of freedom [120]. Because each observation is tested as a potential outlier, the p-values are adjusted for multiplicity via a Benjamini-Hochberg family-wise discovery-rate (FDR) correction [128]. If the adjusted p-value is less than 0.01, the observation is declared an outlier.

Implementing ROUT may be difficult for those who lack sophisticated software that can handle Cauchy regression. We therefore proposed a modification of the ROUT method. Instead of assuming a Cauchy distribution,

robust model fitting is performed via iteratively reweighted least squares (IRWLS) by minimizing $\sum_{j=1}^{n} w_j \{y_j - f(\theta, x_j)\}^2$ using Huber weights with $w_j = \min\left(1, \frac{1.345 \times MAD}{|r_j|}\right)$. Let $SE(r_j) = \sqrt{MAD^2 + \dot{f}\hat{V}\dot{f}^t}$ where $\hat{V}$ is the estimated variance-covariance matrix for $\hat{\theta}$ and $\dot{f} = \frac{\partial f(\partial, x)}{\delta\theta}\Big|_{\theta=\hat{\theta}}$. Note that $SE(r_j)$ is a robust estimate of the standard deviation for a new predicted response evaluated at x. The test is performed by examining the p-value= $2P(T > |r_j|SE(r_j))$, where T follows a central t-distribution with $n-4$ degrees of freedom. As with the ROUT method, the p-values are adjusted for multiplicity via a Benjamini-Hochberg FDR correction. If the adjusted p-value is less than 0.01, the observation is declared an outlier. For abbreviation, given its connection to a prediction interval, this test will be labeled RPI (robust prediction interval).

4.2.2 Tests for Whole Curve Outliers

When the entire concentration-response curve is independently replicated, it is possible that the data produced from one or more of the whole curves is bad. In this work, we consider the case with three replicate curves and attempt to determine if at least one of the whole curves is an outlier. In this section, a second index is added to the response y to indicate the number of curve replicates. It is assumed that $y_{jk} = f(\theta_k, x_j) + \epsilon_{jk}$, with the $q-$length parameter θ_k denoting the model parameters for the k^{th} curve, ϵ_{jk} $\mathcal{N}(0, \sigma^2)$, $k = 1, 2, \ldots K$ (usually $K = 3$) and $j = 1, 2, \ldots, n_k$ (n_k is the number of observation within curve k). In Chapter 3, we define a whole-curve outlier as an entire curve replicate affected by a manipulation error. It can affect the dilution factor (Figure 4.2, left), one of the asymptotes (Figure 4.2, middle), or both asymptotes, inducing a vertical shift (Figure 4.2, right).

Note that, with three replicate curves, should one curve be declared a whole-curve outlier, it may be impossible to judge whether the issue is with the one whole curve or with the other two whole curves. In such a circumstance, one may look to historical data to make a proper judgment or one may simply choose to re-run the experiment. For example, the assay responses of one of the three curves may be historically low-valued, strongly suggesting that it, and not the other two curves, is the whole-curve outlier.

We could not find whole-curve outlier detection methods in the literature.

We propose a novel procedure, which we call the maximum departure test (MDT). This procedure examines the maximum difference across the range of experimental concentrations of the mean response among the three curves and compares this difference to a critical value. The MDT is conducted by first considering a model such as (1) to associate with each replicate curve. Let $X = \{x : x_1 \leq x \leq x_n\}$, where x_1 and x_n denote the minimum and maximum concentrations. Given a pair of curves (k, k'), we permit a small deviation δ so that if $\theta_k \neq \theta_{k'}$, we may declare that the two curves differ if $|f(\theta_k, x) - f(\theta_{k'}, x)| > \delta$ for at least one $x \in X$.

Let $\upsilon_{k,k'} = \max_{x \in X} |f(\theta_k, x) - f(\theta_{k'}, x)|$. Then $\upsilon_{k,k'} > \delta$ indicates that the two curves differ significantly for at least one concentration inside the range of designed concentrations. Instead of $\binom{K}{2}$ separate tests, we propose an omnibus testing parameter given by $\Upsilon_X = \max(\upsilon_{1,2}, \upsilon_{1,3}, \ldots, \upsilon_{1,K}, \ldots, \upsilon_{K-1,K})$. With three curves, the test parameter is $\Upsilon_X = \max(\upsilon_{1,2}, \upsilon_{1,3}, \upsilon_{2,3}$. If $\Upsilon_X > \delta$, at least one curve is out of trend. Formally, the two competing hypotheses are given by

$$\begin{aligned} H_0 &: \Upsilon_X \leq \delta \\ H_a &: \Upsilon_X > \delta \end{aligned} \tag{4.3}$$

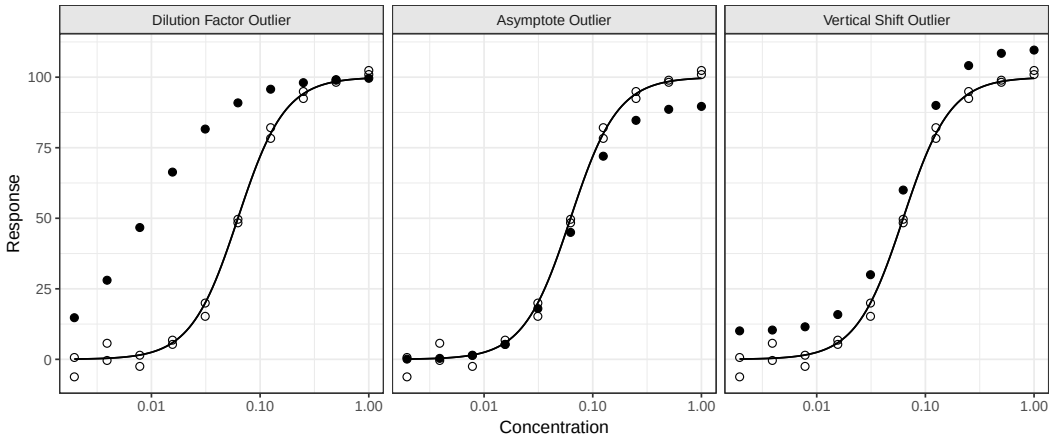


Figure 4.2: Examples of whole curve outliers with an effect on the dilution factor (left), the upper asymptote (middle) or both asymptotes (right)

If H_0 is rejected in Equation 4.3, then at least one whole curve significantly differs from another for one or more concentrations and so, at least one whole curve is declared an outlier. To test Equation 4.3, we propose fitting the three curves separately by least squares, resulting in estimates $(\hat{\theta}_k, \hat{V}_k, \hat{\sigma}_k^2)\,(k = 1, 2, 3)$, where $\hat{\theta}_k$ is the estimator for θ_k, V_k is the estimated variance-covariance matrix for θ_k, and σ_k^2 is the estimated residual variance associated with $n_k - q$ degrees of freedom. A pooled residual variance estimated is

$$\hat{\sigma}^2_{pooled} = \frac{\sum (n_k - q)\,\hat{\sigma}_k^2}{\eta}$$

which is associated with $\eta = \sum (n_k - q)$ degrees of freedom. Let $SE_k(x) =$ standard error of $f(\theta_k, x)$ and $SE_{k'}(x) =$ standard error of $f(\theta_{k'}, x)$. The pooled standard error of $f(\hat{\theta}_k, x) - f(\hat{\theta}_{k'}, x)$ is

$$SE_{k,k'(x)} = \hat{\sigma}_{pooled}\sqrt{\left(\frac{SE_k(x)}{\hat{\sigma}_k}\right)^2 + \left(\frac{SE_{k'}(x)}{\hat{\sigma}_{k'}}\right)^2}$$

If the two curves are exactly δ units apart at concentration x, then asymptotically,

$$\frac{f(\hat{\theta}_k, x) - f(\hat{\theta}_{k'}, x) - \delta}{SE_{k,k'(x)}}$$

follows a central T distribution with η degrees of freedom. It follows that

$$\frac{\left|f(\hat{\theta}_k, x) - f(\hat{\theta}_{k'}, x)\right| - \delta}{SE_{k,k'(x)}}$$

follows a half central T distribution with η degrees of freedom. Let

$$\hat{T}_{k'} = \max_{x \in X} \frac{\left|f(\hat{\theta}_k, x) - f(\hat{\theta}_{k'}, x)\right| - \delta}{SE_{k,k'(x)}}$$

and

$$\hat{T}_X = \max(\hat{T}_{1,2}, \hat{T}_{1,3}, \hat{T}_{2,3})$$

.

Although each of the $\hat{T}_{k'}$ statistics is related to the half central T distribution through order statistics and $\hat{T}_X$ represents a second hierarchical level of order statistics, to keep the algorithm simple, we compare $\hat{T}_X$ to a

half central T distribution with η degrees of freedom, calculating a p-value to test Equation 4.3 with p-value$= 1 - P(T_{1/2} > \hat{T}_X)$, where $T_{1/2}$ is half-T distributed [129]. If $\hat{T}_X < 0$ (i.e., the maximum departure is less than δ), the p-value$= 1$. Because $\hat{T}_X$ is the maximum of three entities, the significance level is adjusted to $\alpha/3$ with $\alpha = 0.01$.

4.3 Computer Simulation

Computer simulation studies were conducted to examine the characteristics of each outlier detection method provided in Section 4.2. In each simulated scenario, except for generated outliers, the reference standard and test curves were created to be identical so that, in lieu of outliers, the curves should be declared similar and the estimated RP should be 100%. We did not generate a computer simulation based on non-similar curves because the similarity testing procedure should filter out those cases whether outliers are present or not. See Chapter 3 for the effect of outliers on similarity testing and estimated RP for non-similar curves. For a given simulation scenario and simulated data set, detected outliers were removed prior to similarity testing and RP estimation. The false positive and true positive outlier-detection rates were estimated as was the similarity-testing failure rate. For reference standard and test curves that were declared similar, the bias in RP was also calculated. Each scenario is simulated 10,000 times.

For each simulated run, three replicates of the reference standard and test samples were computer-generated via model (1) with parameters $\theta_R = \theta_T = (yMin = 0, yMax = 100, C_{50} = 0.0625, S = 2)$, ten concentrations given by $[1, 0.5, 0.25, \ldots, 0.002]$, and a standard deviation of $\sigma = 2$ or $\sigma = 15$. Thus, the reference standard and test curves are each generated with 30 total observations. The reference standard curve was always generated in an outlier-free manner. In addition to a no-outliers scenario, three different outlier scenarios were created for the test samples.

(i) Outliers at specific concentrations: 1, 2, or 3 outliers are generated at a specific concentration. A concentration point outlier occurs when all three observations at a particular concentration are outliers (see Figure 4.1).

(ii) Random outliers: 1, 2, or 3 outliers are generated at possibly different random concentrations.

(iii) Whole curves outliers: Three types of whole-curve outliers are tested separately (see Figure 4.2).

(a) Dilution Factor Outlier

(b) Upper Asymptote Outlier

(c) Vertical Shift Outlier

4.3.1 Single Observation and Concentration Point Outliers

Single observation and concentration point outliers are observed through scenarios (i) and (ii). A factor $k\sigma$ (a multiple of the measurement variability) is added to an observation. Observation outliers can be mild ($|k| = 4$), moderate ($|k| = 6$), or extreme ($|k| = 10$). For scenario (i), outliers were created as all positive ($k > 0$) or all negative ($k < 0$). Because simulation results from all-positive and all-negative outliers were similar, we only show $k > 0$ results for scenario (i). For scenario (ii), outliers can be either positive or negative ($k > 0$ or $k < 0$). Outlier detection was assessed with Hampel's rule, the generalized ESD, the generalized Dixon's Q, ROUT, and the RPI test methods. From the no-outliers scenario, Table 1 shows the proportion of curves (out of 10,000) for which no outliers were detected as well as the proportion for which incorrect declarations of 1, 2, and 3 outliers were made. The false positive rate in Table 4.1 was calculated as the weighted average of incorrect declarations with denominator equal to the number of test sample observations (30). For example, when $\sigma = 2$, the false positive rate for Hampel's test is $(6.5\% \times 1 + 1.3\% \times 2 + 0.4\% \times 3)/30 = 0.3\%$. It appears that, when no outliers are present, all test methods are conservative in terms of test size (i.e., false positive rate $< 1\%$)

For Scenario (i), the true positive rate of each test (ability to correctly identify all outliers) was calculated for each specific concentration of the curve. Figure 4.3 shows the true positive values for moderate outliers. Relative to ROUT and RPI, the USP<1010>-recommended outlier tests (Hampel, ESD, Dixon) performed less well, especially when multiple outliers were

Table 4.1: False positive rate when no outliers are present.

σ	Method	Proportion of curves with # false positive (%)				False positive rate (%)
		0	**1**	**2**	**3+**	
2	Hampel	91.8	6.5	1.3	0.4	0.3
	ESD	99.0	0.9	<0.1	<0.1	<0.1
	Dixon	98.9	1.0	0.1	<0.1	<0.1
	ROUT	88.4	7.9	2.3	1.4	0.6
	RPI	76.2	20.4	3.1	0.4	0.9
15	Hampel	91.4	6.8	1.4	0.4	0.4
	ESD	98.9	0.9	0.1	<0.1	<0.1
	Dixon	98.8	1.1	0.1	<0.1	<0.1
	ROUT	90.8	6.4	1.8	1.0	0.4
	RPI	77.1	19.1	3.3	0.5	0.9

present in the data. The USP<1010> tests and the RPI also experienced trouble detecting outliers in the middle of the curve. Note that, as shown in Chapter 3, outliers in the middle position have the strongest effect on the bias in RP estimation. The pattern of true positive rates is similar for both $\sigma = 2$ and 15. When only one outlier is present, the RPI method is best. When more than one outlier is present at a specific concentration, the ROUT and RPI methods both stand out; however, the ROUT method appears to be more robust to concentration position, as the true positive rate does not drop as low as other methods when outliers are present in the middle of the curve. The mild and extreme outlier cases are similar in behavior to that shown in Figure 4.3, respectively with lower and higher true positive rates.

For Scenario (ii), Figure 4.4 shows the true positive rate of each test, calculated across all concentrations for the situations when 1, 2, and 3 outliers are generated in random positions of the curve and for moderate outliers. As before the ROUT and RPI methods outperform the USP<1010>-proposed test methods. For each test method, mild and extreme outliers were respectively detected at lower and higher true positive rates, as presented in Appendix C. For each outlier severity, results do not appear to depend on σ.

After outlier removal, Equation 4.1 was fitted again to the remaining data

from each sample. Subsequently, each data set was tested for similarity using the same acceptance criteria as for Chapter 3 (see Section 3.5.1). When no outliers are present, the Monte Carlo probability to declare similarity testing was roughly 95% across all outlier detection methods, meaning that the intended lab risk is preserved, regardless of $\sigma = 2$ or 15 and in light of the small number of false positives (see Table 4.1). For scenario (i) and with moderate outliers, Figure 4.5 shows the probability to declare similarity testing after outlier removal for each test method. It is clear that the probability to declare similarity is dependent upon the concentration associated with the outlier. Except for Dixon's method, which worsens in its ability to declare similarity with the magnitude of the outlier (mild/moderate/extreme, see Appendix B for mild and extreme outlier results), the other methods appear invariant to the outlier size. The ROUT method is universally best because

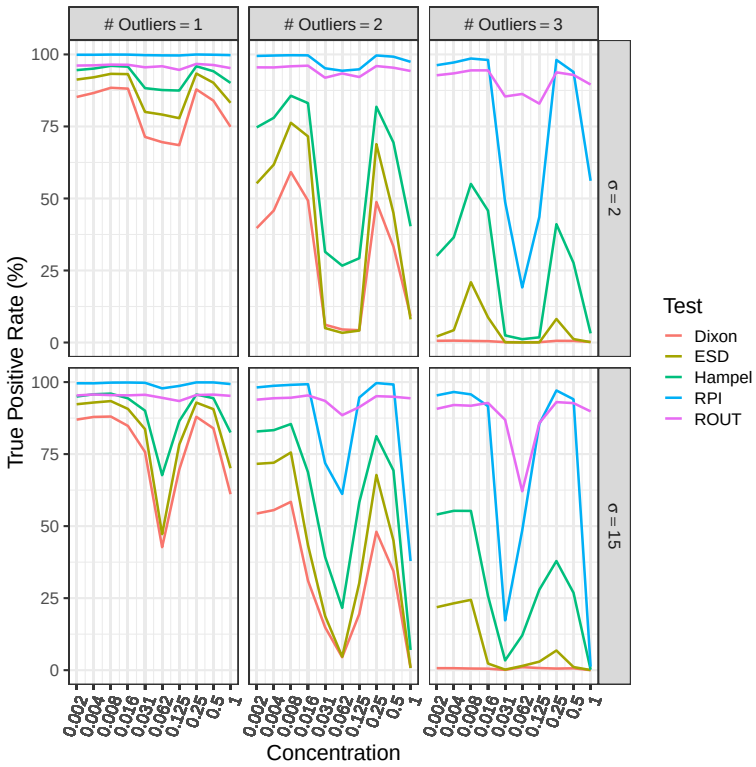


Figure 4.3: True Positive Rate at specific concentration points, moderate outliers

the ability to meet the criteria of the parallelism test is directly linked to the ability to detect and reject and outlier.

We refer to the ratio of the estimated RP to the true RP as recovery. The expected recovery is 100% for each simulation. For scenario (i), Figure 4.6 shows the geometric mean recovery in recovery estimation after outlier removal of moderate outliers for those cases for which similarity is declared, when $\sigma = 2$. Figure 4.7 shows the median coefficient of variation (%CV) of the corresponding recovery (after outlier removal), calculated by $\sqrt{\exp\left(se^2_{\log(RP)}\right) - 1}$, where $se^2_{\log(RP)}$ is the standard error of the estimated log(recovery) from each Monte Carlo run. Based on Figure 4.5, note that the geometric mean recovery for ROUT is calculated on more than 75% of simulated data sets out of 10,000. For the other test methods, the percentage of simulated data sets used to calculate bias in recovery depends highly on concentration, with some test method/concentration pairs approaching 0% out of 10,000. The size/color of the points in Figures 4.6 and 4.6 show correspondence with the percentage of declared similarity cases. ROUT appears to be universally best in minimizing the percent bias and %GCV in recovery. Figure 4.7 suggests that, regardless of correct/incorrect outlier detection, if

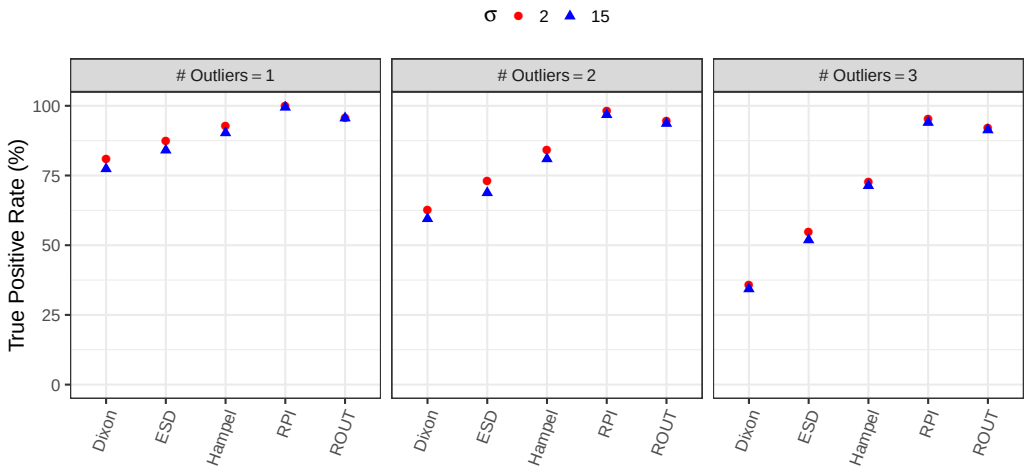


Figure 4.4: True Positive Rate across random concentration points for moderate outliers

the criteria for parallelism are met, the %CV of the ensuing %recovery estimate is relatively stable across methods (%CV $< 4\%$ in all cases). Results for $\sigma = 15$ exhibit the same pattern, with different magnitude. For Hampel, ESD, and Dixon, the percent bias grows larger with increasing outlier magnitude. For RPI and ROUT, the opposite pattern emerges because they detect sensibly all extreme outlier. Results for all outlier severity and σ are presented in Appendix C.

For scenario (ii), Figure 4.8 shows the 95% coverage interval of % recovery in estimated RP when moderate outliers are randomly generated across concentrations. Only the results for which parallelism was declared were taken into account. When no outlier was present, as expected for the intended lab risk, roughly 5% of simulated cases were removed after parallelism testing.

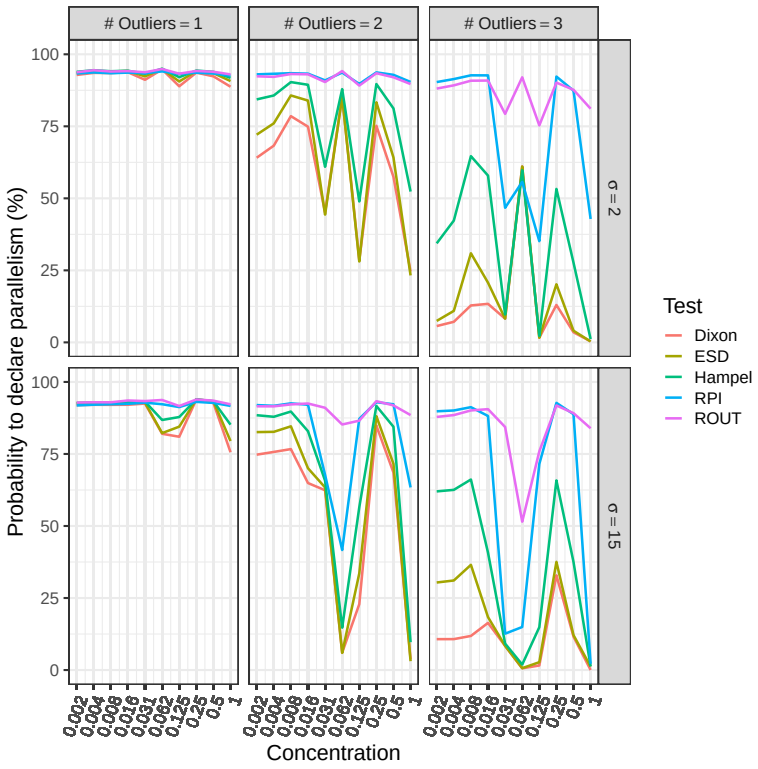


Figure 4.5: Probability to declare similarity test after detection and removal of moderate outliers

When only one outlier was present, we declared parallelism for roughly 90 to 95% of curves for each outlier detection method. When two outliers were present, this percent of cases that met the parallelism testing criteria respectively dropped to about 83-90% or 63-82% for Dixon, ESD and Hampel's tests and for both RPI and ROUT, at least 91% of cases meet the parallelism testing criteria no matter how many outliers were present. When 0 or 1 outliers were present, the predicted recovery cover the same range no matter the detection method. When 2 or 3 outliers were present, ROUT and RPI cover narrower range than the USP<1010> suggested methods.

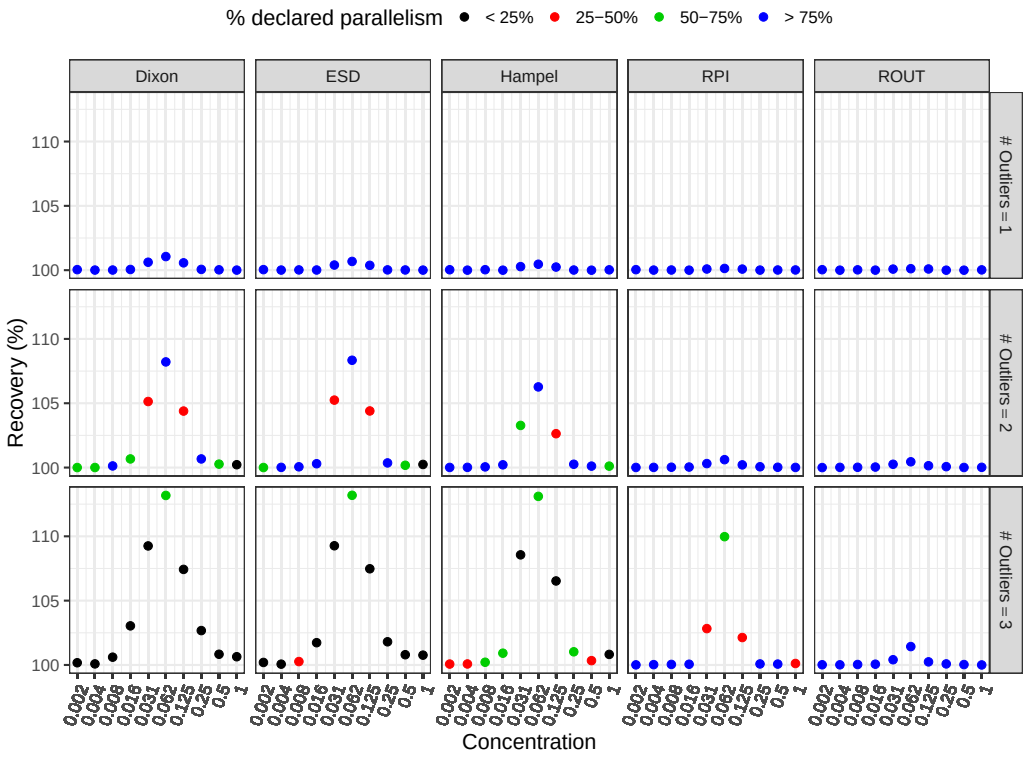


Figure 4.6: Geometric Mean %recovery in estimated relative potency (RP) after detection and removal of moderate outliers when $\sigma = 2$

4.3.2 Whole Curve Outliers

Whole curve outliers are observed through Scenario (iii). Whole curve outliers were generated by changing the parameters:

(a) Dilution Factor Outlier: The dilution factor changes for one of the replicates. Consequently, the true concentration is different from the observed concentration for that curve.

(b) Upper Asymptote Outlier: $yMax$ is different for one of the replicates.

(c) Vertical Shift Outlier: both $yMax$ and $yMin$ are modified the exact same amount for one of the replicates.

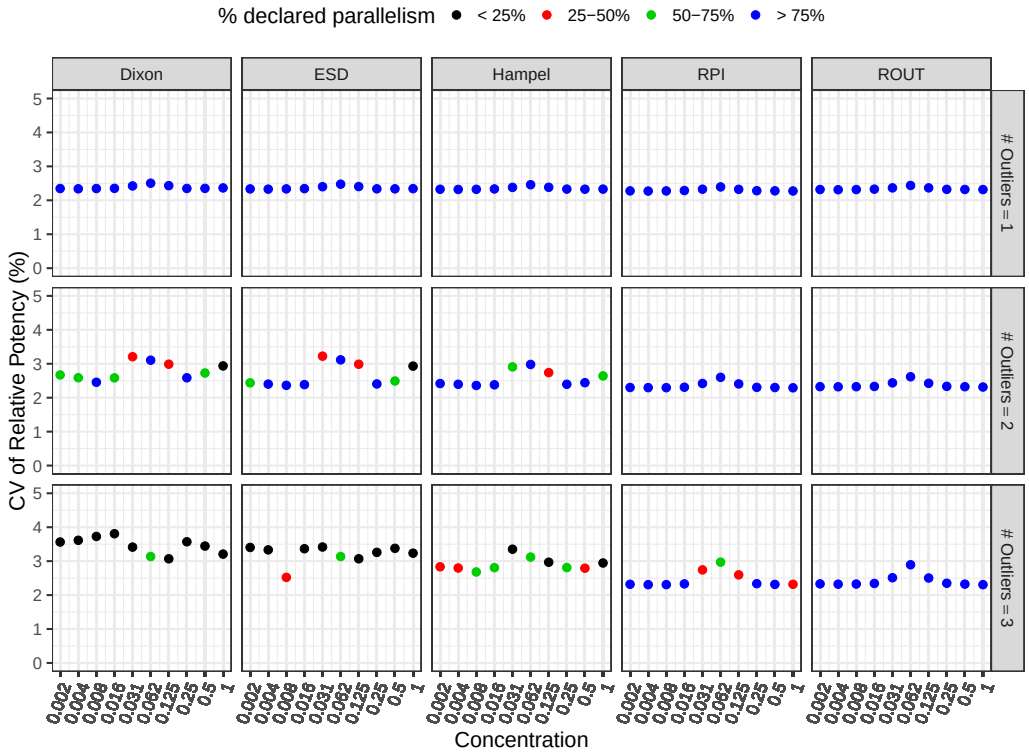


Figure 4.7: Median %CV of the estimated relative potency (RP) after detection and removal of moderate outliers when $\sigma = 2$

Whole-curve outliers can also be mild, moderate, or extreme; as well as positive or negative. Table 4.2 presents the different values for the changed parameters. Each scenario was run 10,000 times with $\sigma = 2$ and $\sigma = 15$.

The proposed MDT method was applied to each simulated data set for whole-curve outlier detection using $\delta = 3.1$. The value of $\delta = 3.1$ was determined through computer simulation to ensure a false positive rate of about 1%. As a naïve comparator, the ROUT method (in our opinion, the best method for single observation and concentration point outlier detection) was applied separately to each curve. If ROUT detected at least three outliers from one or more curves, a whole-curve outlier was declared. We label the ROUT whole-curve outlier method as ROUT3 to indicate its reliance on detecting at least three outliers in one curve. The ROUT3 method, or any

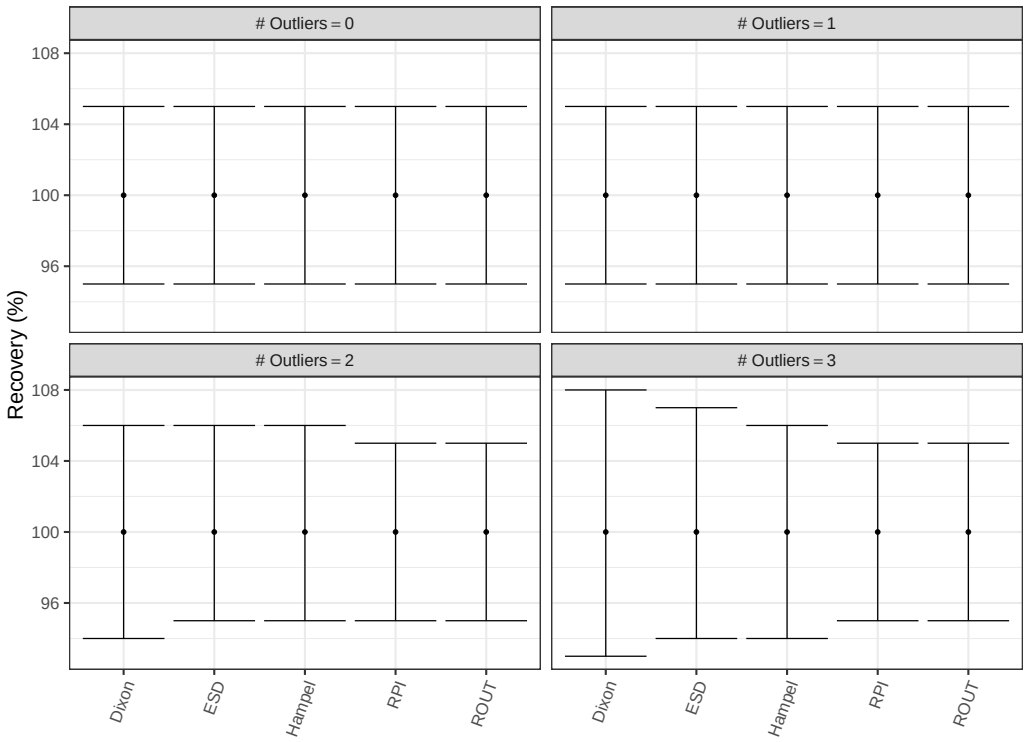


Figure 4.8: 95% Monte Carlo coverage interval of estimated recovery after detection and removal of moderate outliers at random positions $\sigma = 2$

Table 4.2: Values for the whole-curve outliers. The values for the Dilution Factor Outliers are the wrong dilution factors, the values for the Upper Asymptote Outliers is the wrong yMax, and the values for the Vertical Shift Outliers are the delta applied to both yMin and yMax.

	True Value		**Mild**	**Moderate**	**Extreme**
Dilution Factor	01:02.0	Positive	01:01.8	01:01.8	01:01.7
		Negative	01:02.2	01:02.2	01:02.3
Asymptote	100	Positive	98	90	85
		Negative	102	110	115
Vertical Shift	0	Positive	-2	-10	-15
		Negative	2	10	15

of the methods tested in Section 4.2.1 for that matter, is not designed to detect whole-curve outliers and so it is expected that its statistical power to detect whole-curve outliers will be smaller than that of the proposed MDT method. With 3 replicate curves, for the MDT or ROUT3 tests, if a whole-curve outlier is declared, it may be difficult to determine which curve is the outlier and the entire experiment may well be at risk. Thus, we only report the false positive and true positive rates to detect whole-curve outliers.

When no outliers are present, the false positive rate to declare a whole-curve outlier was about 1% for MDT and $< 0.5\%$ for the ROUT3 method. Table 4.3 presents the true-positive rate to detect whole-curve outliers. Neither test seems to be efficient when $\sigma = 15$. This is not surprising considering that severe outliers are only equivalent to a one standard deviation shift. MDT is powerful to detect moderate to severe outliers when $\sigma = 2$. The ROUT3 method performs well in the case of a dilution factor outlier and severe upper asymptote outlier when $\sigma = 2$ but performs very poorly in case of a vertical shift, probably due to the fact that these outliers affect the model uniformly enough for the robust model to fit normally. Overall, the proposed MDT method appears to perform well and should be considered as a reasonable tool for the detection of whole-curve outliers.

Table 4.3: Probability (%) to detect whole curve outliers

σ	Side	Severity	Dilution Factor		Upper Asymptote		Vertical Shift	
			MDT	ROUT3	MDT	ROUT3	MDT	ROUT3
2	Negative	Mild	98	91	<1	<1	<1	<1
		Moderate	99	100	84	45	95	<1
		Extreme	100	100	100	95	100	<1
	Positive	Mild	99	67	<1	<1	<1	<1
		Moderate	100	89	85	45	95	1
		Extreme	100	96	100	95	100	<1
15	Negative	Mild	5	<1	2	<1	2	<1
		Moderate	8	<1	3	<1	5	<1
		Strong	13	1	5	<1	9	<1
	Positive	Mild	3	<1	2	<1	2	<1
		Moderate	4	<1	3	<1	4	<1
		Strong	5	<1	6	<1	9	<1

4.4 Discussion

In the case of potency bioassays, USP<1032> recommends screening for outliers. In Chapter 3, we showed the effect of multiple types of outlier on parallelism testing and Relative Potency estimation, confirming the need to detect those outliers. USP<1010> proposes several tests for outlier detection. This Chapter showed that the USP<1010>-proposed tests are out-performed by robust regression outlier detection methods (ROUT/RPI) in terms of true-positive outlier detection, declaration of similarity, and bias in the RP value. It is our opinion that the ROUT method is best for single observation and concentration point outliers in potency bioassays. The RPI method also performed well and might best be used by those who have no access to a Cauchy maximum likelihood algorithm.

For whole-curve outliers, we proposed the Maximum Departure Test (MDT) and compared it to the ROUT3 method. The MDT method fared well in detecting whole-curve outliers. Future research will focus on competing whole-curve outlier testing.

Our findings suggest that robust regressions, such as the Cauchy regression used for the ROUT method, could be an efficient way to fit different nonlinear models, related or not related to potency assays. Future work will investigate a generalized approach.

A limitation of our study is that it assumes that the residuals are independent and normally distributed, and homogeneity of variance across the dilution profile. Future work should investigate the consequence of the violation of one or more of these assumptions.

This page is intentionally left blank

Chapter 5

Efficient Design to Estimate Relative Potency and Test for Similarity in Parallel Curve Assays

This Chapter is based on a manuscript called "Simple Efficient Design to Estimate Relative Potency and Test for Similarity in Parallel Curve Assays", soon to be submitted for publication in a scientific journal.

5.1 Introduction

In relative potency (RP) assays, the RP can be determined by computing the horizontal difference between the log(concentration)-response functions of the test and reference [47]. The computed value is only relevant if the concentration-response functions of the two products (or batches, lots, samples), are similar [34]. In the case of parallel curve assays, this similarity is demonstrated statistically, by showing that, in the log(concentration) scale, the function of the test product is a horizontal shift from the reference standard's function [47]. If the functions are not parallel, the horizontal difference is not constant (Figure 1.3).

To model the concentration-response function of a preparation, the literature commonly suggests the use of a four-parameter logistic (4PL) model. The most common of these was proposed by Rodbard and Hutt [50], described in Section 1.2.1:

$$y_{ij} = yMax_i + \frac{yMin_i - yMax_i}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^{S_i}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{5.1}$$

While mathematically equivalent, the following parametrization presents computational advantages [51]:

$$y_{ij} = yMax_i + \frac{yMin_i - yMax_i}{1 + \exp\left(S_i \left(\log(x_{ij}) + c_i\right)\right)} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{5.2}$$

where $c_i = \log(C_{50_i})$. We note $\theta_i = [yMin_i, yMax_i, c_i, S_i]^t$.

Parallelism is accepted if the lower asymptotes, upper asymptotes and growth rates are similar between the two curves (Figure 1.3, right). If parallelism is declared, the relative potency is then the estimated horizontal distance between the two curves on a log scale after fitting the model

$$y_{ij} = yMax + \frac{yMin - yMax}{1 + \exp\left(S \left(\log(x_{ij}) + c_i\right)\right)} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{5.3}$$

Which is an adaptation of Equation 5.2 with common $yMin$, $yMax$, and S for both curves. The horizontal distance between the two curves is then the difference between c_R and c_T and the RP is calculated by $\exp(\rho)$ with $\rho = c_R - c_T$. Multiple tests for similarity have been suggested in the literature [79, 84, 88, 90]. A common challenge in laboratories is selecting the best concentration points — or, more generally, support points — to estimate the relative potency a well as the non-similarity metric(s) of choice.

Multiple options for the optimal-design to estimate the parameters of linear models are proposed in the literature [130–132]. A common optimality criterion for both linear and nonlinear models is the D-optimality, which

maximizes the determinant of the information matrix and consequently minimizes the joint confidence region of the parameters. Criteria also exist to find the best design to estimate the ratio of two linear combinations of the vector of parameters, which is used for the inverse prediction — useful to calculate assay potency — as well as estimating slope ratios for parallelism assessment in parallel line assays [133, 134]. For nonlinear models, such as the 4PL, Mukkula and Paulen (2017) showed that the linearization-based approach fails to identify the optimal xs [135]. Locally optimal designs for nonlinear models were first introduced by Chernoff [136]. Melas (2004) later demonstrated that the support points of locally D-optimal designs are functions of the nonlinear parameters [137]. Kalish and Rosenberger (1978) first derived a D-optimal design for two-parameter logistics curves [138], and Bezeau and Endrenyi (1986) extended it to three-parameter logistics models [139]. Finding the D-optimal design in linear model is trivial through matrix algebra. For nonlinear model, computational tools, such as an imperialist competitive algorithm are needed [140]. However, finding the D-optimal design may not be of interest, or even realistic, in the case of relative potency assays. Another known issue with classic locally optimal design is that they depend on the true value of the model parameters being known. This problem is usually addressed using Bayesian optimal designs [141]. We did not address this specific issue in the present work, and it will be the focus of future research. However, the true parameter may be impossible to know because it varies from one sample to another. Khinkis et al. (2003) claim that a D-optimal design for logistic models remains optimal as long as it is re-evaluated when the true value of the parameters shifts [142]. In the case of potency assays, one or more curve parameters may vary from one plate to another, and their true value for a specific plate cannot be known in advance. Additionally, an optimal design does not account for laboratory constraints in the lab. Support points often cannot be decided at liberty. In serial dilution assays, each evaluated concentration is a dilution of the previous one. Consequently, in the log scale, points are usually equally spaced across the concentration range. Khinkis et al. refer to this as log-spread designs and claim that locally D-optimal are more efficient but realize that they may not always be practical when making serial dilutions in 96-well plates [142]. A limitation of their study is that they compare designs using D-optimality as a performance criterion and were therefore likely to find that D-optimal design is better. François et al. propose to use optimality criteria that are more relevant to estimate a potency [143]. However, they focus on single signal back

prediction over a calibration curves and therefore ignore the non-similarity criteria, which also need to be considered for potency assays.

Another aspect, not yet accounted for in the literature, is that two curves need to be modeled, with prior information that might be available for the reference product if it has been used in the past on similar set up. The relative potency as well as the non-similarity parameters for each new test sample are unknown until measured. In this work, we therefore advocate that numerical optimality — such as D, A, G, I , or E optimality — should not be the objective. Instead, we suggest finding an efficient and practical design with support points that will allow reasonably accurate and reasonably precise estimation of the RP and non-similarity metrics. In Section 5.2, we describe and perform a simulation to propose a methodology to select concentration points, with only estimates of the reference curve parameters, while accounting for lab constrains. In Section 5.3, we assess the robustness of our approach by changing the reference curve parameters, with and without a correct prior estimate of the parameter. Finally, in Section 5.4, we discuss the present work and propose future research.

5.2 Efficient Designs

5.2.1 What Operators Can Control

When designing the serial dilution, scientists control three aspects: the range of concentrations (1); the number of equally spaced log(concentration)-points across this range, noted n_c (2); and the number of replicates at each concentration for each curve, noted n_{rep} (3). Although not mandatory, in most assays, number of replicates is the same at every concentration. It is also related to lowering the overall uncertainty around parameter estimates rather than to the optimal concentration-points. Therefore, while an important aspect in practice, n_{rep} is not considered in this paper to find an efficient design. Other aspects, like separating experiments into blocks (runs, plates, . . .), can be considered. Obviously, those aspects are laboratory — and often assay — specific, and therefore are not considered in this manuscript but may be the subject of later research.

The ideal range of concentrations depends on the curve parameters, and

more specifically depends on the C_{50}. We note ζ the range of log(concentrations) that is necessary to observe responses ranging from 0.5% to 99.5% of the distance between the lower and upper asymptote of the reference curve, and we suggest increasing the width of ζ by a factor κ to find the most efficient design. The log(concentration) range becomes

$$Z = \bar{\zeta} \pm \frac{\kappa \times \Delta\zeta}{2} \tag{5.4}$$

For example, if $\theta_R = [yMin_R = 0, yMax_R = 1, c_R = \log(0.0625), 2]^t$, then

$$\zeta = \log\left(\left[g\left(\theta_R, 0.005\right); g\left(\theta_R, 0.995\right)\right]\right) = [-5.419; -0.126]$$

With $g(\theta_R, y)$ the inverse function of the 4PL. Then, if $\kappa = 2$, the log(concentrations) range becomes

$$Z = -2.773 \pm \frac{2 \times 5.293}{2} = [-8.066; 2.520]$$

The challenge is to find the ideal factor κ. Later in this section, we propose to use computer simulation to do so, using the information available for a specific assay. Once Z is determined, the log(dilution factor) depends on the number of concentration-points and the log(dilution-factor), or the spacing of log(concentration)-points, is then determined by $\frac{\Delta Z}{n_c - 1}$.

5.2.2 Non-Similarity and Relative Potency

For a design to be efficient, it must accurately and precisely estimate the value of interest that are used during assay routine use: first, the non-similarity metrics using model (2) and second, for curves that are declared similar, the relative potency using model (3). Many tests for similarity have been proposed in the literature. In this work, we focus on three tests: the ratio of curve parameters [84], the residual sum of squared error due to non-parallelism [79], and the maximum departure test [88, 90].

Yang et al. (2012) proposed the following ratios of parameters:

$$r_1 = \frac{yMax_T}{yMax_R}$$
$$r_2 = \frac{yMax_T - yMin_T}{yMax_R - yMin_R}$$
$$r_3 = \frac{(yMax_T - yMin_T)S_T}{(yMax_R - yMin_R)S_R}$$

r_3 represents the ratio of the slopes at the inflection point. We use log transformed ratios to ensure symmetry, as suggested by Berger and Hsu (1996) [86], and note $\lambda_k = \log(r_k)$.

Gottschalk and Dunn (2005) proposed $RSSE_{nonpar} = RSSE_{constrained} - RSSE_{full}$ as a metric for non-similarity, where $RSSE_{full}$ and $RSSE_{constrained}$ are the residual sum of squared calculated after fitting respectively models 5.2 and 5.3 on the pair of curves [79]. $RSSE_{nonpar}$ is sometimes referred to as χ^2-metric [118].

Novick et al. (2012) suggested to use the metric

$$\xi = \min_{RP} \max_{x_L, x_U} \left| CI_{95}\left(f\left(x, \hat{\theta}_R\right) - f\left(x \times RP, \hat{\theta}_T\right)\right)\right|$$

where x_L and x_U are the ranges of concentrations across which parallelism is assessed, $f\left(\hat{\theta}_R, x\right)$ is the original least square (OLS) estimate of the 4PL curve fit of the reference product, $f\left(\hat{\theta}_T, x\right)$ is the OLS estimate of 4PL curve fit of the test product, and $CI_{95}(h)$ is a notation for the 95% confidence interval of h [88]. The original methodology used Bayesian methods to estimate ξ. Novick and Yang (2019) later proposed a frequentist approximation [90]. We used the latter in this Chapter.

5.2.3 Simulation Setup

We want to find a concentration range width multiplication factor κ (see Equation 5.4), combined with a concentration-points spacing, that would consistently lead to good estimations. The quality of these estimation also depends on the measurement variability, which cannot be decreased by design, except by increasing the number of replicates per concentration. Those

two aspects are not considered in the simulations. Additionally, we assume constant variability across the concentration range, and no variation in the curve parameters associated with blocking (or other structural element of the assay design) or dilution. In practice, if these assumptions are not verified, some changes to the design may be necessary.

We generated 1,500 designs. We set $yMin_R = 0$, and the mean value of c_R at log(0.0625). We draw a different value for $c_R \sim \mathcal{N}\left(log\left(0.0625\right), \sigma_c^2\right)$ for each set of curves within one design. The 1,500 designs were generated using a Latin-hypercube space-filling design for computer experiments [144]. The simulation parameters and associated ranges of possible values are presented in Table 5.1.

Table 5.1: Ranges for each simulation parameter

Simulation Parameter	**Range of possible values**
Steepness of reference curve (S_R)	[1.5–4]
Upper asymptote of reference curve ($yMax_R$)	[1–3]
Log(Relative Potency) (ρ)	[log(0.25)–log(4)]
Log(ratio) of upper asymptotes (λ_1)	[log(0.8)–log(1.25)]
Log(ratio) of asymptote ranges (λ_2)	[log(0.8)–log(1.25)]
Log(ratio) of slopes (λ_3)	[log(1/1.5)–log(1.5)]
Range width multiplication factor (κ)	[1–3]
Number of concentrations points (n_c)	[6–12]
Between-plates variability of the log(C_{50_R}) (σ_c^2)	$\log(1.1)^2$–$\log(2)^2$]

For each design, we generated 1,000 plates each containing three curves: a reference, and two samples. We calculate ζ and Z from κ and the curve parameters of the reference, using 0.0625 for the C_{50}, because the true value changes for each plate and is always unknown.

The true values for the first sample curve ($T1$) parameters θ_{T1} are determined by:

- $yMin_{T1} = r_1 yMax_R - r_2\left(yMax_R - yMin_R\right)$
- $yMax_{T1} = r_1 yMax_R$
- $C_{50_{T1}} = \frac{C_{50_R}}{RP}$

- $S_{T1} = S_R \frac{r_3}{r_2}$

The second sample ($T2$) curve is exactly parallel to the reference curve. The true values for θ_{T2} are therefore exactly equal to θ_R except for $C_{50_{T2}} = \frac{C_{50_R}}{RP}$.

The non-similarity metrics discussed in this paper are estimated by comparing the reference and $T1$. Estimating the RP on non-similar curves is irrelevant and determining acceptance limits for each of these metrics is out of scope for this manuscript. We therefore estimated the RP only from fitting Equation 5.3 on the reference curve and $T2$.

Optimal designs for nonlinear models only depend on the prior estimation of each parameters, and not on the residual variance [145]. To observe that effect, we duplicated every design and generated curves with $\sigma = 2\%$ or 10% of $(yMax_R - yMin_R)$. Because we fixed $yMin_R = 0$, this is equivalent to $\sigma = 2\%$ or 10% of $yMax_R$.

5.2.4 Simulation Findings

For the relative potency and curve parameter ratios, we considered the 95^{th} percentile of the absolute difference between the observed statistics and the design-specific true values. For $RSSE_{nonpar}$ and ξ, no true value exists, so we analyzed the Monte Carlo standard deviation of the estimated test statistics. We only presented here the results for $\sigma = 2\%$ of $yMax_R$. Results for $\sigma = 10\%$ of $yMax_R$ present a similar pattern (with obviously a different magnitude) and are presented in Appendix D.1.

Figure 5.1 shows the 95^{th} percentile of the absolute difference between the observed and true log(relative potency) as a function of the true log(relative potency) ρ, the range width multiplication factor κ, and the number of concentration points n_c. The error in relative potency estimation seems to increase when n_c decreases. Specifically, high errors appear when $n_c \leq 7$ if the range is large ($\kappa > 2$) and the relative potency is high or low ($|\rho| > .5$). Other simulation parameters did not seem to affect the relative potency estimation error. The 95^{th} percentile of the absolute difference between the observed and true log(relative potency) as a function of each of the design specific parameters is presented in Appendix .

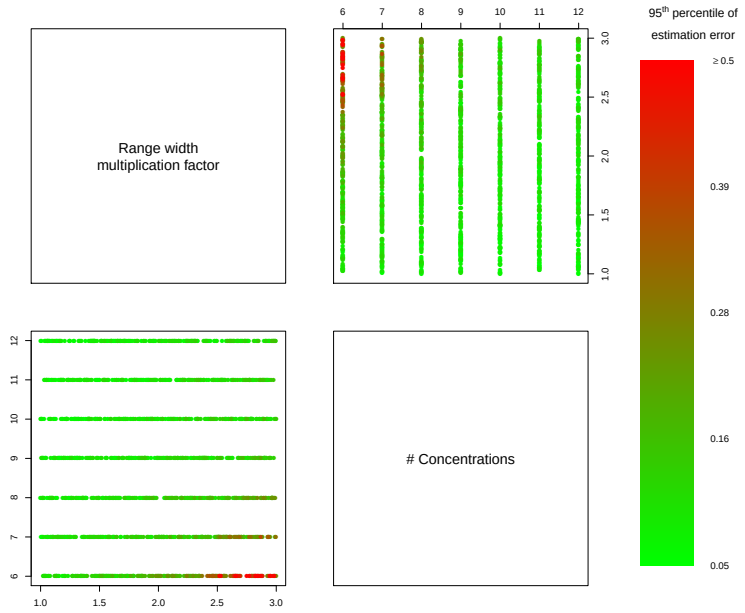


Figure 5.1: Scatterplot matrix of the 95^{th} percentile of the absolute difference between the observed and true log(RP) as a function of the log(RP), range width multiplication factor and number of concentrations when $\sigma = 0.02 \times yMax_R$. Light green points represent simulation designs with small estimation error and red points represent simulation designs with high estimation error.

Figures 5.2 and 5.3 present the 95^{th} percentile of the absolute difference between the observed and true log(ratios) of upper asymptotes λ_1 and ranges between asymptotes λ_2 as a function of each of the log(relative potency), the range width multiplication factor κ, and the steepness of the reference curve. Figure 5.4 present the log(ratio) of slopes λ_3 as a function of κ and the number of concentrations n_c. The error in estimation error of λ_1, and λ_2 is high when the concentration range is narrow ($\kappa < 2$), the curve is steep ($S_R > 3$), and the relative potency is high. Low relative potencies also affect λ_2, as both asymptotes need to be observed for an accurate and precise observation of the asymptote range. Inversely, the estimation error of λ_3 becomes high when the range is too large, especially if $n_c < 10$. Other simulation parameters did not seem to affect the curve parameter ratios estimation error. The 95^{th} percentiles of the absolute difference between the observed and true log(ratios) as a function of each of the design specific

parameters are presented in Appendix .

No clear design pattern could be identified for the $RSSE_{nonpar}$ and maximum departure ξ, so no result is presented in the main manuscript for these two non-similarity metrics. The Monte Carlo standard deviations of $RSSE_{nonpar}$ and ξ as a function of each of the design specific parameters are presented in Appendix .

Overall, this simulation results suggest that a concentration range width multiplication factor of $\kappa = 2$ combined with a number of concentration points $n_c \geq 10$ should ensure an efficient estimation of the RP and non-similarity metrics, relative to the measurement variability.

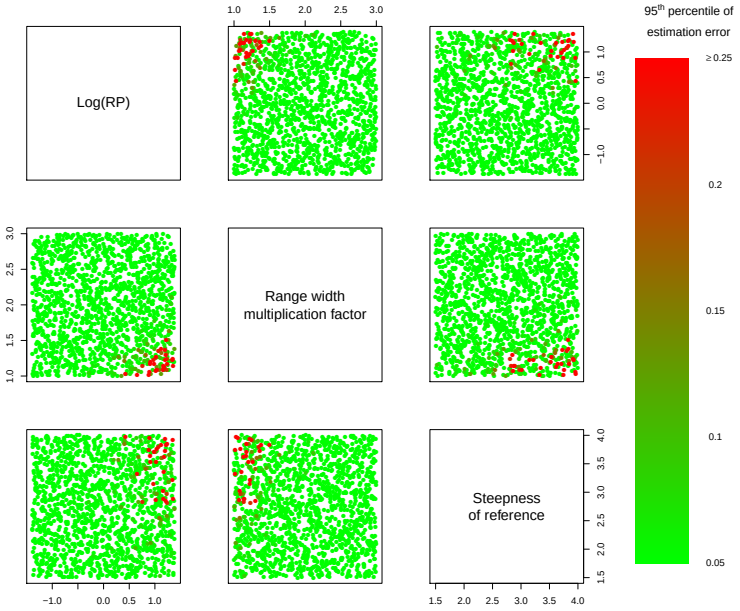


Figure 5.2: Scatterplot matrix of the 95^{th} percentile of the absolute difference between the observed and true log(ratio) of upper asymptotes as a function of the log(RP), range width multiplication factor and steepness of reference curve when $\sigma = 0.02 \times yMax_R$. Light green points represent simulation designs with small estimation error and red points represent simulation designs with high estimation error.

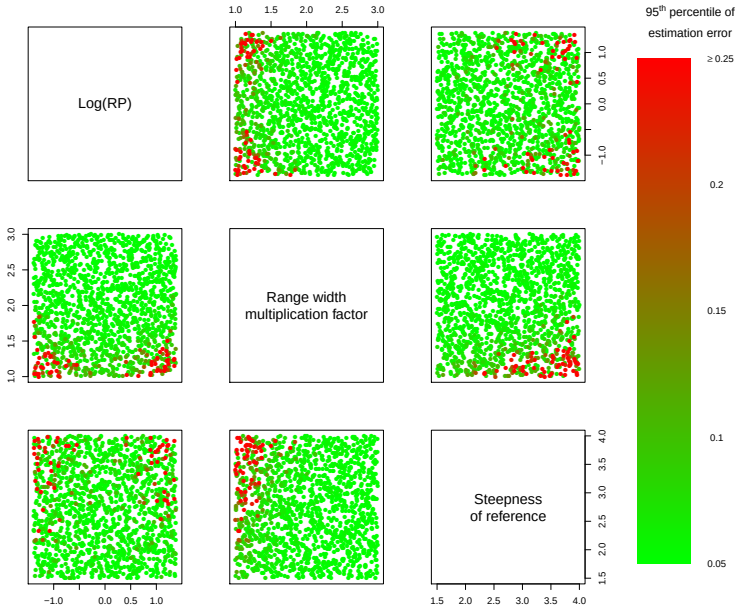


Figure 5.3: Scatterplot matrix of the 95^{th} percentile of the absolute difference between the observed and true log(ratio) of asymptote ranges as a function of the log(RP), range width multiplication factor and steepness of reference curve when $\sigma = 0.02 \times yMax_R$. Light green points represent simulation designs with small estimation error and red points represent simulation designs with high estimation error.

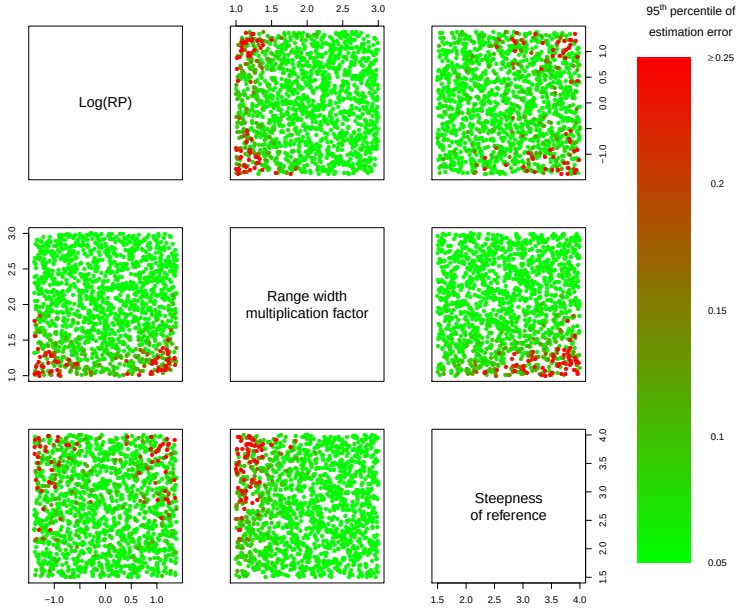


Figure 5.4: Scatterplot matrix of the 95^{th} percentile of the absolute difference between the observed and true log(ratio) of slopes as a function of the range width multiplication factor and number of concentrations when $\sigma = 0.02 \times yMax_R$. Light green points represent simulation designs with small estimation error and red points represent simulation designs with high estimation error.

5.3 Robustness assessment

From the simulations presented in Section 5.2, we observed that the true values of $yMin$, $yMax$, and c do not affect the quality of parameter estimation if the concentration range is calculated using Equation 5.4. However, like D-optimal designs, Equation 5.4 relies on the knowledge of θ_R. We note θ_R^a the assumed values for θ_R. All analysis from Section 5.2 assume that $\theta_R^a = \theta_R$. We ran a second simulation study to assess the robustness of using $\kappa = 2$ and $n_c = 10$ when$\theta_R^a \neq \theta_R$. As steeper curves and high relative potencies lead to worse results, we always consider $S_R^a = 4$ and $RP = 400\%$ in order to test for the "worst-case" scenario. We varied $yMax_R$, $yMin_R$, and c_R while fixing $\theta_R^a = [0, 1, \log(0.0625), 4]$. A total of 125 scenarios were analyzed, with every combination of the following simulation parameters:

- $yMax_R = 1 \times [0.5, 0.8, 1, 1.25, 2]$
- $c_R = \log(0.0625) \times [0.5, 0.8, 1, 1.25, 2]$
- $S_R = 4 \times [0.5, 0.8, 1, 1.25, 2]$

For all scenarios, we used n_{rep} and $\sigma = 0.02 \times yMax_R$. From each scenario, we generated 1,000 plates each containing two curves: a reference, and one sample. Because the ratios of non-C_{50} curve parameters did not seem to affect the quality of estimations, we always consider $r_1 = r_2 = r_3 = 1$ for simplicity. The true values for θ_T are therefore exactly equal to θ_R except $C_{50_T} = \frac{C_{50_R}}{RP}$.

Because θ_R^a is fixed, the range of log(concentrations) to observe 99% of the reference response range is the same for every scenario and is calculated by

$$\zeta = \log\left(\left[g\left(\theta_R^a, 0.005\right); g\left(\theta_R^a, 0.995\right)\right]\right) = [-4.096; -1.449]$$

Then, as $\kappa = 2$, the log(concentration) range becomes

$$Z = -2.773 \pm \frac{2 \times 2.647}{2} = [-5.419; -0.126]$$

For the robustness study, we focused only on the quality of estimation of the log(relative potency) and log(ratio) of non-C_{50} parameters.

Figure 5.5 presents the 95^{th} percentile of the absolute difference between the observed and true value of the log(RP) and log (ratio) of the upper asymptote, the range between asymptotes, and the slopes as a function of the under (or over) prior assumption of the reference curve parameters. From our simulations, it appears that highly over or under-assuming the C_{50} of the reference (respectively $c_R = 0.5 \times c_R^a$ or $c_R = 2 \times c_R^a$) has negative effects on the estimation of all parameters of interest. Additionally, over-assuming the reference curve steepness ($S_R < S_R^a$) increases the error in log(relative potency) estimation; while under-assuming the steepness ($S_R > S_R^a$) increases the error in estimation of the log(ratio) of slopes and increases the negative effect of poor C_{50} assumption. Under-assumption of the reference curve upper asymptote ($yMax_R > yMax_R^a$) seems to increase the negative effect of the reference steepness under-assumption on the log(ratio) of slopes. Overall a range width multiplication factor of 2 with 10 concentrations is robust to reference curve parameters assumed within 80 to 125% of the true values.

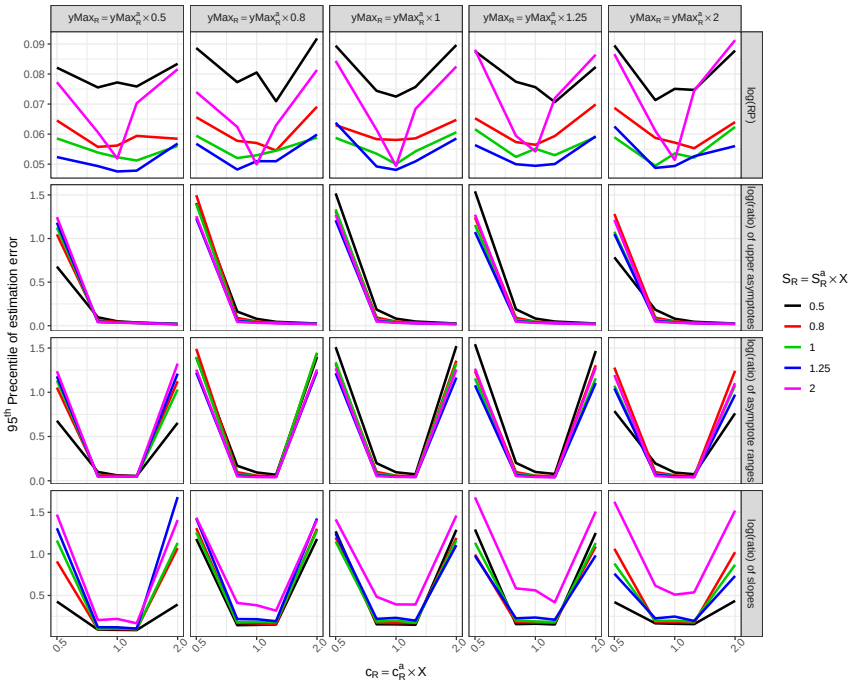


Figure 5.5: 95^{th} percentile of the absolute difference between the observed and true value of the log(RP) and log(ratio) of the upper asymptote, the range between asymptotes, and the slopes as a function of the under (or over) prior assumption of the reference curve parameters.

5.4 Discussion

Optimal designs for nonlinear model fail to account for the laboratory constraints and actual objectives. Tusto et al (2016) proposed D-Optimal strategies specifically for the relative potency estimation, but they focused on the two-parameter logistic curves, so they did not have to worry about demonstrating asymptote similarity nor observing both asymptotes for each curve [146]. In this paper, we propose a simple way to find efficient concentration-points for relative potency estimation and similarity testing. A range that is wide enough to observe 99% of the response range of the reference preparation is too narrow to estimate low/high relative potencies. We therefore suggested to double the width of this range, and log-spread 10 concentrations. This multiplication factor may need to be adapted if the between-plate coef-

ficient variation in C_{50} is greater than 100% ($\sigma_c > \log(2)$), or if the expected relative potency range falls outside 25-400%.

This methodology is robust to a certain level of wrong parameter estimations, but still requires a solid prior knowledge of the reference curve parameter values. Future research on this topic will focus on applying Bayesian theory to this method, in order to account for the uncertainty around the prior estimates.

Another limitation of our proposed approach is that, when the curve is too steep, the found concentrations range may become too narrow for both asymptotes to be seen in case of high/low relative potencies. In case of highly steep curve (e.g. $S_R > 4$), a possible solution is to use only 8 concentrations within the calculated range, then add one level of dilution on each side.

Chapter 6

Comparison of Parallelism Tests for Potency Assays

This Chapter is based on the article titled "Extensive Comparison of Parallelism Tests for Potency Assays", soon to be submitted for publication in a scientific journal. As of writing, we are waiting for cross-referenced papers to be published before submitting.

6.1 Introduction

In potency assays, the relative potency (RP) is defined as the horizontal difference between the log(concentration)-response functions of a test and a reference products, usually depicted by logistic curves function [47]. The RP computed value is relevant when the two curves are parallel (also referred to as similar). Indeed, lack of parallelism between the two curves indicates non-similarity in the biological activity, or binding affinity, within the assays for the test and the reference products [34]. Furthermore, by definition, parallelism indicates that the function of the test product is a horizontal shift from the reference standard's function [84]. Therefore, if the functions are not parallel, the horizontal difference is not equal to a single constant value over the observed concentration range (Figure 1.3).

The most used model for concentration-response functions is the four-parameter logistic (4PL) [147], and a common parametrization of this model

was proposed by Rodbard and Hutt [50], described in Section 1.2.1:

$$y_{ij} = yMax_i + \frac{yMin_i - yMax_i}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^{S_i}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{6.1}$$

We note $\theta_i = [yMin_i, yMax_i, C_{50_i}, S_i]^t$.

Functions are parallel if they share the same lower asymptotes, upper asymptotes and effect rates. In case parallelism is declared, the RP can be calculated by the ratio of C_{50}s after fitting the model:

$$y_{ij} = yMax + \frac{yMin - yMax}{1 + \left(\frac{x_{ij}}{C_{50_i}}\right)^{S}} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{6.2}$$

Equation 6.2 is an adaptation of Equation 6.1 with common $yMin$, $yMax$ and S for both curves.

Historically, to test for parallelism, the residuals obtained from fitting Equations 6.2 and 6.1 were commonly compared by an $F - Ratio$ [148, 149]. This $F - Ratio$ compares the lack of fit of the constrained model due to non-parallelism to its lack of fit due to other reasons. The test statistic is computed as follows:

$$\frac{(RSSE_{constrained} - RSSE_{full})\,(df_{constrained} - df_{full})}{(RSSE_{full})\,(df_{full})} \sim F_{(df_{constrained} - df_{full}, df_{full})} \tag{6.3}$$

where $RSSE_{full}$ and $RSSE_{constrained}$ are respectively the residual sums of squares (RSSE) of the models obtained by fitting Equations 6.1 and 6.2, respectively, and df_{full} and $df_{constrained}$ are their associated degrees of freedom.

The $F - Ratio$ test is known to increase the chances of passing parallelism for curves with high variability. In order to correct that weakness, Gottschalk and Dunn (2005) proposed an alternative approach based on the difference between the weighted RSSEs (wRSSE) of the models from Equations 6.2 and 6.1 [79] [79]:

$$wRSSE_{constrained} - wRSSE_{full} \sim \chi^2_{df_{constrained} - df_{full}} \tag{6.4}$$

A drawback of this χ^2-test is that it requires weighted regressions, which is not recommended for serial dilution models [80, 81]. Assuming that a weighing is actually needed, the correct weights are unknown and tend to be poorly estimated. Additionally, like the $F - Ratio$ test, this test declares parallelism in case of lack of statistical significance. This is fundamentally flawed, and hypotheses are now considered such that statistical significance is required to demonstrate parallelism [82–84].

Multiple equivalence tests for parallelism have been proposed in the literature. Each test is claimed to be the correct way to address parallelism, and their advantages are presented using limited examples. This manuscript aims to extensively compare the performances of multiple proposed tests using a large simulation study. In Section 6.2, we introduce the tests that are considered in the present study, as well as the simulation setup. In Section 6.3, the simulation results are presented. In Section 6.4, we discuss the different results and comment on the evaluated tests.

6.2 Material and Method

This Section presents the different tests that are evaluated in this study as well as the simulation plan.

6.2.1 Proposed Parallelism Tests

6.2.1.1 Empirical $RSSE_{nonpar}test$

In their original form, the difference tests ($F - Ratio$ and χ^2) have been discussed in the literature as not appropriate to assess parallelism [82,83,150]. As a consequence, these original versions were not considered in this study. An alternative to calculating a $p - value$ based on a χ^2 distribution is to compare the reference to itself a large number of times. A percentile of the obtained non-weighted $RSSE_{nonpar}$s is then declared as the acceptance

limit [118, 151]. According to this alternative, two curves are declared to be parallel if

$$RSSE_{nonpar} \leq RSSE_{crit} \tag{6.5}$$

where $RSSE_{crit}$ is the calculated percentile.

6.2.1.2 Hyper-Rectangle test with confidence intervals (HR_{ci})

The second test considered in this paper was proposed by Yang et al. (2012). It suggests to separately fit the 4PL model to both reference and test curves and then compare the confidence intervals of the estimated ratio of the upper asymptotes (r_1), the ranges between the asymptotes (r_2), and the slopes at the inflection point (r_3) to pre-defined equivalence margins [47].

$$\begin{aligned} r_1 &= \frac{yMax_T}{yMax_R} \\ r_2 &= \frac{yMax_T - yMin_T}{yMax_R - yMin_R} \\ r_3 &= \frac{(yMax_T - yMin_T)S_T}{(yMax_R - yMin_R)S_R} \end{aligned}$$

Parallelism is declared if

$$D_{Lk} \leq CI_{90}(r_k) \leq D_{Uk} \forall k \in 1, 2, 3 \tag{6.6}$$

where D_{Lk} and D_{Uk} are respectively the lower and upper acceptance limits for $CI_{90}(r_k)$ and $CI_{90}(r_k)$ is the 90% confidence interval for r_k. While presenting the ratios on a 3-dimension scale, the acceptance region within limits D_{Lk} and D_{Uk} represent a hyper-rectangle. In this study, we work with the log-ratios as suggested by Berger and Hsu (1996) [86].

6.2.1.3 Hyper-Rectangle test without confidence intervals (HR_{noci})

By definition, confidence intervals are used when a decision for a population is made using a sample drawn from that population. Parallelism tests are

often used to accept or reject a relative potency sample. In that case, confidence interval is irrelevant because the decision is made on the sample, using that same sample. The Hyper-Rectangle test without confidence intervals becomes

$$d_{Lk} \leq \hat{r}_k \leq d_{Uk} \forall k \in 1, 2, 3 \tag{6.7}$$

where d_{Lk} and d_{Uk} are respectively the lower and upper acceptance limits for r_k.

6.2.1.4 Ellipsoid test

In Chapter 2, we proposed to replace the hyper-rectangle by an ellipsoid. Assuming that the estimate of the log(ratios) under true parallelism follow a three-dimension multivariate normal distribution centered in $[0, 0, 0]^t$ with variance-covariance matrix Σ, a future pair of curves is declared parallel if it falls within the derived ellipsoid:

$$\hat{\lambda}^t \hat{\Sigma} \hat{\lambda} \leq q_{\alpha,3} \tag{6.8}$$

where $\hat{\lambda} = \log([\hat{r}_1, \hat{r}_2, \hat{r}_3]^t)$; and $q_{\alpha,3}$ is the $100(1-\alpha)^{th}$ quantile of a χ^2 distribution with 3 degrees of freedom.

6.2.1.5 Maximum Departure Test (MDT)

Novick et al. (2012) claim that analyzing ratios of parameters is not enough to declare similarity between the entire nonlinear profiles [88, 89]. Instead, they suggest looking at the maximum departure between the confidence intervals of the reference curve fit and the test curve fit after the optimal horizontal shift. They first used a Bayesian method to calculate the confidence intervals [88] and later proposed a frequentist approximation [90]. Parallelism is declared if

$$\min_{RP} \max_{x_L, x_U} \left| CI_{95} \left(f\left(x, \hat{\theta}_R\right) - f\left(x \times RP, \hat{\theta}_T\right) \right) \right| \leq \delta \tag{6.9}$$

where x_L and x_U are the ranges of concentrations across which parallelism is assessed, $f\left(\hat{\theta}_R, x\right)$ is the original least square (OLS) estimate of the 4PL curve fit of the reference product, $f\left(\hat{\theta}_T, x\right)$ is the OLS estimate of 4PL curve fit of the test product, and δ is the maximum accepted departure from parallelism to declare similarity. In this study, we used the frequentist approximation.

6.2.2 Simulation Setup

Computer simulation studies were conducted to explore the characteristics of each parallelism test presented in Section 6.2.1. For various scenario described hereafter:

- $\theta_R = [0, 1, 0.0625, 2]^t$
- $100,000$ pairs of curves were generated with $\theta_T = \theta_R$
- $100,000$ pairs of curves were generated with $\theta_T \neq \theta_R$
- y was generated for both the reference (y_R) and test (y_T) at 10 concentration, with three replicates per concentration.

4 different designs were considered at 2 different measurement variabilities. In addition to the 'normal' case, 2 stressed situations were proposed. A total of 24 scenarios is evaluated:

- 4 designs (see Figure 6.1)
  - Design 1, full view of the curve: $x = [12.5, 3.85, 1.18, \ldots, 0.0003]$
  - Design 2, full view of the curve: $x = [0.25, 0.125, 0.0625, \ldots, 0.0003]$
  - Design 3, full view of the curve: $x = [100, 12.5, 1.5625, \ldots, 0.0000007]$
  - Design 4, full view of the curve: $x = [0.25, 0.1786, 0.1276, \ldots, 0.0121]$
- 2 measurement variabilities
  - $\sigma = 0.02; 0.05$

- Stressed scenarios:
  - Unstressed, everything stays normal after margins were calculated.
  - Presence of outliers: 1 concentration point (randomly selected for each simulation) was assumed to have been declared as outliers and therefore removed in the test curve.
  - Variability increase: we assumed that, after validation and derivation of acceptance margins, the measurement variability had increased 50%. Therefore, σ becomes 0.03 or 0.075.

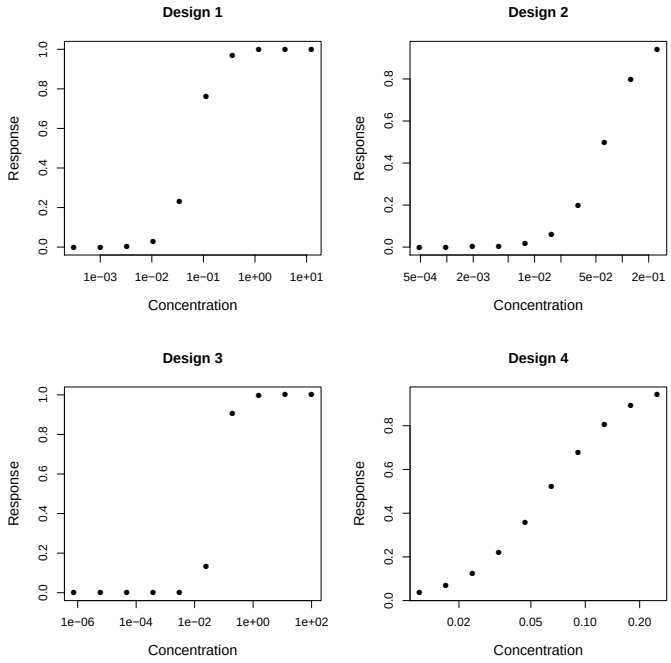


Figure 6.1: Representation of a reference curve in each of the assessed designs

USP<1032> suggests observing reference to reference comparisons to establish acceptance limits. Acceptance limits were calculated for each test, separately by design [47]. The distribution of each statistical test statistic under true parallelism was obtained from the 100,000 pairs of curves generated with $\theta_T = \theta_R$ in the unstressed scenario. We calculated margins to aim

a lab risk, noted α, of 10%, 5%, and 1%. This is possible because, in this simulations study, we know the true θ_R. If θ_R is unknown, MCMC simulations can be performed to observe reference to reference comparisons (see Chapter 2) [152, 153].

To generate curves with $\theta_T \neq \theta_R$, we randomly chose r_1, r_2, r_2 and RP from the following distributions:

- $r_1 = \exp\left(\mathcal{U}\left(\log\left(0.8\right), \log\left(1.25\right)\right)\right)$
- $r_2 = \exp\left(\mathcal{U}\left(\log\left(0.8\right), \log\left(1.25\right)\right)\right)$
- $r_3 = \exp\left(\mathcal{U}\left(\log\left(1/1.5\right), \log\left(1.5\right)\right)\right)$
- $RP = \exp\left(\mathcal{U}\left(\log\left(0.5\right), \log\left(2.0\right)\right)\right)$

With $\mathcal{U}(a, b)$ a Uniform distribution between a and b. These distributions were chosen to ensure that the full range of accepted λs would be contained in the simulations.

Let $\theta_T = [yMin_T, yMax_T, C_{50_T}, S_T]^t$, with

- $yMin_T = r_1 yMax_R - r_2\left(yMax_R - yMin_R\right)$
- $yMax_T = r_1 yMax_R$
- $C_{50_T} = \frac{C_{50_R}}{RP}$
- $S_T = S_R \frac{r_3}{r_2}$

Quantifying the consumer risk as a percentage of non-similar curves accepted was impossible because an infinity of non-similar curves scenarios exists. Instead we considered the consumer risk in terms of the relative error in relative potency, noted β. For each pair of curves that passes a specific similarity test, we fit Equation 6.2 and estimate the RP. Then calculate the absolute relative % error:

$$\psi = 100 \times \left| \frac{\text{Observed RP} - \text{True RP}}{\text{True RP}} \right| \tag{6.10}$$

We calculated ψ for each of the 100,000 curves of each scenario and defined the consumer risk β as the 95^{th} percentile of ψ for each considered statistical test and each α. We then compared the lab and consumer risks of each test (see Section 6.3).

6.3 Simulation Results

6.3.1 Calculation of Acceptance Limits

Acceptance limits were calculated for each considered statistical test, separately by design and measurement variability. For the empirical $RSSE_{nonpar}$ and maximum departure tests, the $100(1-\alpha)^{th}$ percentile of the test statistics distribution under true parallelism were used as acceptance limits (see Table 6.1). Novick and Yang (2019) suggest using a limit of $\delta = 5\sigma$ for a lab risk of approximatively 1% [90]. With three replicates, $5 \times 0.02/\sqrt{3} = 0.0577$ and $5 \times 0.05/\sqrt{3} = 0.1443$. According to the results presented in Table 6.1, our findings seem aligned with this claim if the design is appropriate. For designs 3 and 4, this claims not verified.

Table 6.1: Acceptance limits for $RSSE_{nonpar}$ and MDT by design, σ and α

Design	σ	**RSSE$_{\text{nonpar}}$**			**MDT**		
		$\alpha = \mathbf{0.10}$	$\alpha = \mathbf{0.05}$	$\alpha = \mathbf{0.01}$	$\alpha = \mathbf{0.10}$	$\alpha = \mathbf{0.05}$	$\alpha = \mathbf{0.01}$
1	0.02	0.0005	0.0008	0.0012	0.0440	0.0484	0.0570
1	0.05	0.0052	0.0065	0.0093	0.1162	0.1278	0.1505
2	0.02	0.0006	0.0008	0.0012	0.0436	0.0474	0.0546
2	0.05	0.0052	0.0065	0.0094	0.1121	0.1211	0.1383
3	0.02	0.0005	0.0007	0.0012	0.1221	0.1847	0.5539
3	0.05	0.0049	0.0062	0.0090	0.8653	0.9640	1.1650
4	0.02	0.0005	0.0008	0.0012	0.1419	0.1618	0.2155
4	0.05	0.0052	0.0065	0.0095	0.3662	0.4309	0.5449

The smallest hyper-rectangles containing $100(1-\alpha)\%$ of the joint distribution of $CI_{90}(\lambda)$ or λ under true parallelism were derived using Nelder and Mead's optimization method to obtain the acceptance limits for the hyper

rectangle tests with and without 90% confidence intervals, respectively (see Table 6.2) [61]. From Table 6.2, we observe that the limits for HR_{ci} are sometimes so large for λ_3 — up to 0.6540 for design 1, and up to 12.9295 for design 3, respectively corresponding to $1.92^{\pm 1}$ and $412,297^{\pm 1}$ in the original scale — that they lose all biological relevance.

For the ellipsoid, the Pearson covariance matrix of the 100,000 $\hat{\lambda}$s under true parallelism was calculated for each design and each measurement variability. Equation 6.11 shows the different $\hat{\Sigma}_{design,\sigma}$, with design=1,2, 3, or 4 and $\sigma = 0.02$ or 0.05.

$$
\begin{aligned}
\hat{\Sigma}_{1,0.02} &= \begin{bmatrix} 0.00007 & 0.00007 & -0.00007 \\ 0.00007 & 0.00017 & -0.00017 \\ -0.00007 & -0.00017 & 0.00250 \end{bmatrix} \\
\hat{\Sigma}_{1,0.05} &= \begin{bmatrix} 0.00043 & 0.00047 & -0.00049 \\ 0.00047 & 0.00108 & -0.00126 \\ -0.00049 & -0.00126 & 0.01762 \end{bmatrix} \\
\hat{\Sigma}_{2,0.02} &= \begin{bmatrix} 0.00045 & 0.00049 & -0.00042 \\ 0.00049 & 0.00059 & -0.00054 \\ -0.00042 & -0.00054 & 0.00216 \end{bmatrix} \\
\hat{\Sigma}_{2,0.05} &= \begin{bmatrix} 0.00324 & 0.00349 & -0.00312 \\ 0.00349 & 0.00417 & -0.00395 \\ -0.00312 & -0.00395 & 0.01504 \end{bmatrix} \\
\hat{\Sigma}_{3,0.02} &= \begin{bmatrix} 0.00008 & 0.00009 & -0.00035 \\ 0.00009 & 0.00015 & -0.00067 \\ -0.00035 & -0.00067 & 0.04319 \end{bmatrix} \\
\hat{\Sigma}_{3,0.05} &= \begin{bmatrix} 0.00052 & 0.00054 & -0.00200 \\ 0.00054 & 0.00091 & -0.00361 \\ -0.00200 & -0.00361 & 0.22067 \end{bmatrix} \\
\hat{\Sigma}_{4,0.02} &= \begin{bmatrix} 0.00043 & 0.00089 & -0.00049 \\ 0.00089 & 0.00488 & -0.00106 \\ -0.00049 & -0.00106 & 0.00176 \end{bmatrix} \\
\hat{\Sigma}_{4,0.05} &= \begin{bmatrix} 0.00340 & 0.00782 & -0.00314 \\ 0.00782 & 0.05444 & -0.00192 \\ -0.00314 & -0.00192 & 0.01616 \end{bmatrix}
\end{aligned}
\tag{6.11}
$$

Table 6.2: Acceptance limits for the confidence interval of each log(ratio) by design, σ and α. Because the acceptance limits are symmetrical around zero, only the positive value is presented.

Design	σ	α	$\mathbf{HR_{ci}}$			$\mathbf{HR_{noci}}$		
			$\mathbf{CI_{90}}(\lambda_1)$	$\mathbf{CI_{90}}(\lambda_2)$	$\mathbf{CI_{90}}(\lambda_3)$	λ_1	λ_2	λ_3
1	0.02	0.10	0.0341	0.0515	0.2015	0.0230	0.0250	0.1024
1	0.02	0.05	0.0346	0.0529	0.2508	0.0227	0.0288	0.1203
1	0.02	0.01	0.0398	0.0741	0.2483	0.0279	0.0355	0.1549
1	0.05	0.10	0.0915	0.1242	0.5181	0.0421	0.0630	0.3084
1	0.05	0.05	0.0836	0.1502	0.5912	0.0474	0.0762	0.3308
1	0.05	0.01	0.1018	0.1640	0.6540	0.0625	0.0944	0.3885
2	0.02	0.10	0.0825	0.0972	0.1936	0.0413	0.0440	0.1093
2	0.02	0.05	0.0905	0.1065	0.2088	0.0454	0.0546	0.1251
2	0.02	0.01	0.1040	0.1239	0.2549	0.0618	0.0728	0.1407
2	0.05	0.10	0.2206	0.2478	0.5084	0.1109	0.1160	0.2697
2	0.05	0.05	0.2444	0.2816	0.5530	0.1268	0.1384	0.3072
2	0.05	0.01	0.3224	0.3373	0.6530	0.1667	0.1867	0.3762
3	0.02	0.10	0.0450	0.0573	0.9878	0.0209	0.0287	0.3444
3	0.02	0.05	0.0529	0.0692	1.9666	0.0246	0.0326	0.5804
3	0.02	0.01	0.0578	0.0830	5.0992	0.0267	0.0334	1.0096
3	0.05	0.10	0.0813	0.1418	12.022	0.0398	0.0658	1.1564
3	0.05	0.05	0.0908	0.1341	12.813	0.0527	0.0660	1.1572
3	0.05	0.01	0.1428	0.1533	12.930	0.0611	0.0910	1.3335
4	0.02	0.10	0.0874	0.3096	0.1740	0.0432	0.1458	0.0893
4	0.02	0.05	0.0955	0.3393	0.1950	0.0484	0.1832	0.1019
4	0.02	0.01	0.1122	0.4476	0.2168	0.0657	0.2227	0.1280
4	0.05	0.10	0.2760	1.0849	0.5104	0.1377	0.3960	0.2516
4	0.05	0.05	0.3264	1.4526	0.5838	0.1406	0.5786	0.3028
4	0.05	0.01	0.5079	2.6854	1.3338	0.2223	0.9633	0.4084

6.3.2 Evaluation of Consumer and Lab Risks

The consumer risk is the 95^{th} percentile of the error in relative potency for each statistical test in each scenario. The lab risk is the proportion of pairs of curves that failed similarity tests when $\theta_T = \theta_R$. The plots of the consumer risk against the lab risk for each test, design and stress are shown in Figure 6.2 and Figure 6.3 for $\sigma = 0.02$ and 0.05, respectively. The ellipsoid test generally has the lowest consumer risk in the unstressed scenario when $\sigma = 0.02$, and the ellipsoid test and $RSSE_{nonpar}$ seem to be equivalent when $\sigma = 0.05$, except for design 4. All tests increase their lab and consumer risk when outliers are removed from the analysis or when the variability increases. $RSSE_{nonpar}$ loses more power to detect non parallelism in case of outlier removal than the ellipsoid but rejects more curves when the measurement variability increases over time. The MDT seems to be most affected by poor design, as its consumer risk for designs 3 and 4 are high compared to other tests. For designs 1 and 2, the hyper-rectangle tests, with or without confidence intervals, generally have the highest consumer risks.

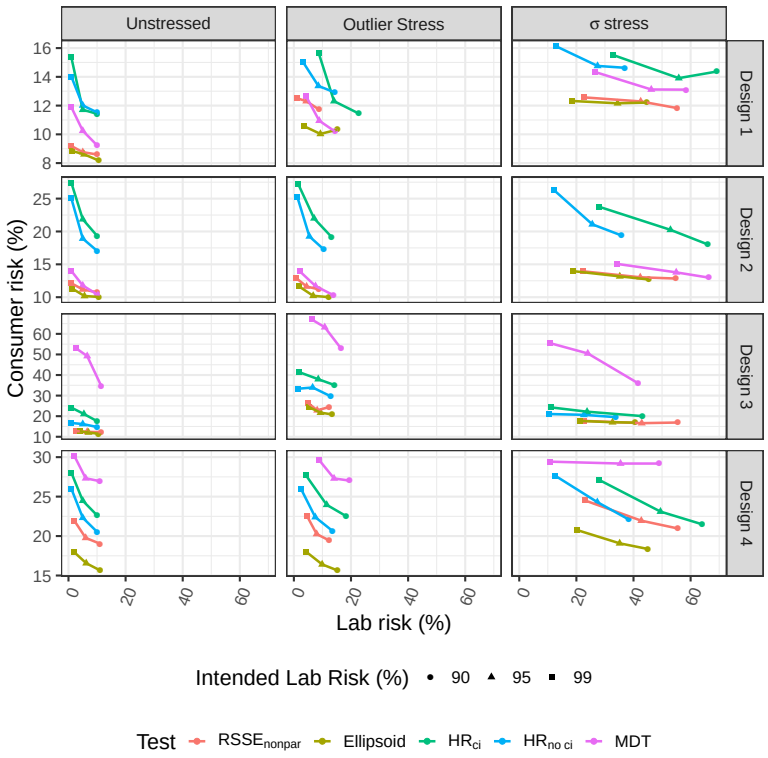


Figure 6.2: Consumer risk vs lab risk by test, design and stress when $\sigma = 0.02$

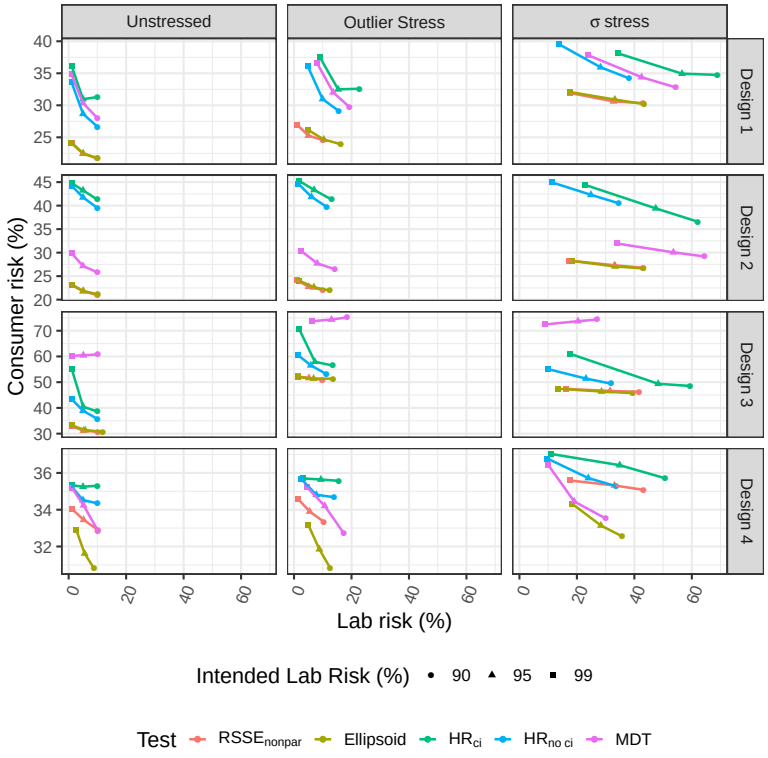


Figure 6.3: Consumer risk vs lab risk by test, design and stress when $\sigma = 0.05$

6.4 Discussion

Many tests for parallelism have been proposed in the literature. The advantages of each of these tests are presented using limited examples. This paper proposes an extensive comparison of popular similarity tests for potency assays, across multiple scenarios. The lab risk was calculated as the proportion of truly parallel curves that fail parallelism test. The consumer risk was calculated as the 95^{th} percentile of the relative error in potency estimation for curves that pass similarity.

The results show that the hyper-rectangle tests have generally high-consumer risks. In addition, the limits for the ratios of parameters are so high (up to $\exp(\pm 12.9295)$ or $\exp(\pm 1.3335)$, respectively with or without confidence intervals) that they lose all biological relevance. We therefore recommend moving away from these tests. The MDT presents acceptable results for the unstressed scenario as long as the design is suitable for the assay purpose. Novick and Yang (2019) suggest using a limit of $\delta = 5\sigma$ for a lab risk of approximatively 1% and use the residual variability of the reference curve as an estimate for σ [90]. This gives the MDT a solid advantage as it does not require calculation of the limit separately for each assay. When the design is poor, however, the MDT doesn't perform well, so this test should only be performed when the concentration can be chosen accordingly.

The ellipsoid and $RSSE_{nonpar}$ tests generally perform better than the other tests, with smaller consumer risks and similar lab risks. After calculation of the acceptance limit by an expert statistician, the $RSSE_{nonpar}$ has the advantage to be directly calculated in most statistical and assay software, as it was introduced in the early 2000s as a difference test [79]. However, the ellipsoid test seems to perform better in most cases and is easily implemented. In addition, while it assesses similarity from all three ratios simultaneously, it still encourages to report and monitor r_1, r_2 and r_3 separately, as they all need to be calculated to perform the test. This can be particularly useful in case of an investigation to understand why an assay repeatedly fails similarity, for example.

Lastly, a limitation of the present study is that all tests have been evaluated with the assumption of homoscedasticity. Intuitively, it seems that composite measures such as the maximum departure and $RSSE_{nonpar}$ would be more affected by heteroscedasticity than parameter ratios estimate. How-

ever, verifying this claim falls outside the scope of this paper and will be assessed in future research.

Chapter 7

General Discussion and Conclusion

7.1 General Discussion

Potency is measured throughout the entire pharmaceutical product development process. A classic measure for the potency of a substance is the C_{50}, which is the concentration needed to reach half of the response range of said substance. However, because the use of the C_{50} as the measure of potency is not always possible, the potency is often measured relative to a reference preparation. The relative potency (RP) is the horizontal distance between the log(concentration)-response curve of the reference and tested preparations. For the RP to be a unique value, this distance must be constant through the entire dilution profile. It is therefore necessary to demonstrate the parallelism between the test and reference preparation concentration-response functions before calculating the RP [47, 53]. This parallelism is called statistical similarity. Historically, similarity was only assessed from late clinical phases forward. However, since the publication of the Quality by Design (QbD) concepts in ICH-Q8 in 2009 [93] for the development of pharmaceutical products, there has been an increasing agreement that the same paradigm also applies to LBAs and bioassays. The validation and routine use of an assay should be of primary importance from the moment it is developed, and similarity (or parallelism) measures should always be considered as critical quality attributes.

Similarity testing was historically performed using p-value based tests, or difference tests. The main flaw of these approaches was that failure to prove non-parallelism between a pair of log(concentration)-response curves resulted in acceptance of similarity. Lack of statistical significance was considered as sufficient proof of equivalence. This is now disregarded in the literature and equivalence tests prevail. The principal challenge that arose with this new paradigm of equivalency was the requirement of a defined "zone of declared similarity". USP <1032> [47] proposes repeated comparisons of the reference product to itself to identify the magnitude of difference when the two products are exactly similar. These comparisons can control the risk of rejecting similarity in case of true parallelism (lab risk) but fail to control the risk of accepting curves which result in high error in relative potency estimation (consumer risk). In Chapter 2, we proposed a three-step derivation of the zone of declared similarity to control for both consumer and lab risk, even when reference-to-reference comparisons are not available. These steps heavily rely on Markov chain Monte Carlo (MCMC) methods, which makes them unappealing to practitioners with little to no experience in Bayesian statistics. However, as Bayesian statistics are becoming increasingly common among non-clinical biostatisticians [154], our approach should be amenable to an increasing number of users. We also proposed a multivariate similarity test, that simultaneously compares all the parallelism-related parameters of the reference and test curves. This is an improvement to the marginal assessment of each parameter which ignores the correlation between the different test statistics.

USP<1032> recommends screening for outliers prior to testing for similarity and calculating the RP. However, the effect of an outlier on the quality of the RP estimate had never been quantified to confirm the utility of outlier removal. Chapter 3 examined the effect on the performance of parallelism testing and the magnitude of the error in RP estimation induced by the presence of a statistical outlier for various outlier types that may be expected in practice, including whole curve outliers. Based on our observations, it appears that outliers close to the C_{50} have little effect on similarity assessment, but sometimes lead to high error in RP estimation. Conversely, outliers in the asymptote do not affect the RP estimation, but may lead to falsely reject parallel curves or falsely accept non-parallel curves. USP<1010> proposes several tests for outlier detection. However, these tests focus on outliers in reported RPs, rather than on outliers that affect the 4PL model. An outlier

in reportable results should be investigated before any action is taken — as rejecting a reportable observation solely on the basis of its relative magnitude may not be advisable [155] — while an outlier that affects the curve fit should always be removed. The performance of the USP<1010>-proposed tests on sigmoid curves was described in Chapter 4. Their performance is inferior to robust regression outlier detection methods (ROUT/RPI) in terms of true-positive outlier detection. Better detection leads to better declaration of similarity (or non-similarity), and smaller error in the reported RP value. Table 7.1 summarizes the pros and cons of each compared outlier test for single observations. For whole-curve outliers, which can affect both aspects, no test was proposed in the literature. We proposed a Maximum Departure Test, which adequately detected whole-curve outliers.

Table 7.1: Summary of pros and cons of each outlier test for single observations

Test	**Pros**	**Cons**
Dixon	Directly available in many software, recommended by USP, very low false positive rate.	Overall lowest detection rate, designed for one outlier only.
ESD	Directly available in many software, recommended by USP, very low false positive rate.	Very low detection rate, mostly for multiple outliers.
Hampel	Easily implemented, recommended by USP, non-parametric.	Very low detection rate, fixed decision limit without rationale for critical value adjustment.
ROUT	Overall best performing test.	Not implemented in CFR 21 compliant commercial software [156]. Complex methodology to implement and validate.
RPI	Performs almost as well as ROUT in many cases, while being easier to implement.	Not currently implemented in any software.

Another challenging aspect of potency assays is the choice of ideal con-

centrations for the concentration-response curve analysis. Optimal designs for nonlinear models fail to account for the laboratory constraints and actual objectives beyond the quality of the fit. In Chapter 5, we proposed a simple way to identify efficient concentration-points for relative potency estimation and similarity testing. While we do not identify the D-Optimal design, our proposed methodology permits an efficient estimation of the relative potency and non-similarity parameters.

Many tests for parallelism have been proposed in the literature. The advantages of each of these tests are often presented using limited examples. In Chapter 6, we compared the parallelism tests discussed in Section 1.2.5, as well as the novel test proposed in Chapter 2. We assessed their performances in an ideal scenario, using the design proposed in Chapter 5; as well as stressed situations, such as post-removal of outliers and the use of sub-optimal concentration support points. The ellipsoid and $RSSE_{nonpar}$ tests generally perform better than the other tests, with smaller consumer risk and similar lab risk. After calculation of the acceptance limit by an expert statistician, the $RSSE_{nonpar}$ has the significant advantage direct calculation possible with most statistical and assay software. However, the ellipsoid test proposed in Chapter 2 seems to perform better in most cases and is easily implemented after estimation of the zone of declared similarity. Note that these tests have been compared as sample suitability tests. If rather than measuring the RP, the assay has the specific objective to assess similarity between two products, tests including confidence intervals should be used. The ellipsoid test can be extended to this purpose but this extension is outside the scope of the present work. Table 7.1 summarizes the pros and cons of each compared parallelism test.

Table 7.2: Summary of pros and cons of each outlier test for single observations

Test	**Pros**	**Cons**
HR_{ci}	Implemented in some validated software, easy to interpret, recommended by USP.	Very high consumer risk, biologically irrelevant acceptance limits in order to control lab risk.
HR_{noci}	Implemented in some validated software, easiest to apply and interpret.	Very high consumer risk.
MDT	No need to pre-define acceptance limits.	Relies heavily on a very efficient design, not trivial to implement.
Ellipsoid	Overall best performing test in observed scenarios.	Most complex test for which to derive acceptance criteria.
$RSSE_{nonpar}$	Implemented as a difference test in many validated software, performs almost as well as the Ellipsoid test in i.i.d. cases.	Probably highly dependent on homogeneity of variance (to be investigated).

7.2 Future work

A limitation of all chapters is that all computations are performed with the assumption of homoscedasticity. A common way to address heteroscedasticity to fit the 4PL model is the use of weighted least square. This is typically a poor choice in bioassay models, as the non-equal variance can usually be linked to a location (or other) effect; thus, nonlinear mixed models are preferred [80,81]. Nevertheless, every aspect of this dissertation can be extended to the heteroscedastic case. The zone of declared similarity derivation proposed in Chapter 2 should work similarly with or without additional variance components. Further investigation could evaluate the effect on the shape of the ellipsoid, as well as the log-normality assumption requested for proper performance of our proposed similarity test. Additionally, the best outlier detection methods and best similarity tests may not be the same in homoscedastic and heteroscedastic systems. Currently, our claim on best methodology only apply to the former, and future work will evaluate the latter.

The ellipsoid test proposed in Chapter 2 requires the use of Bayesian statistics to derive the ellipsoid for every new assay. This expertise is necessary to derive the variance-covariance matrix of the posterior distribution of the log(ratios) of curve parameters in case of true parallelism (Σ). The need for the expertise makes the methodology less widely feasible than other tests. However, the Delta method, and other frequentist approximations, do not provide the correct matrix. Future research may focus on an acceptable approximation for Σ that relies less on MCMC simulations. Alternatively, a computer-based solution to automate the estimation of Σ as proposed could be developed. Another future enhancement could be the evaluation of the effect of prior update on the acceptance limits. As more data become available, the prior knowledge on the reference curve parameters and their different sources of variability increases, and the zone of declared similarity should be periodically reevaluated to reflect that knowledge.

Future research could also change the order of risk control by inverting the last two steps of the zone of declared similarity derivation proposed in Chapter 2. Currently, the first step computes the posterior distribution of the test statistics, the second step derives acceptance limits while accounting for lab risk, and the third step controls for consumer risk. It would be interesting to present an alternative order, and build acceptance limits based

on the consumer risk, then control for lab risk or perform a power study to achieve a desired lab risk.

Robust regression methods have been shown to be useful for outlier detection but are currently only used for this exact purpose. Assessing similarity and estimating the RP directly using robust regression to bypass outlier detection and removal could also be the topic of future studies. Furthermore, our proposed method for whole-curve outliers is adequate, but does not perform as well as outlier detection methods for single observations. As each curve represents one serial dilution, outlier detection methods for functional data could be considered in future work [157, 158].

The methodology proposed in Chapter 5 is robust to some level of parameter estimations inaccuracy, but, like D-Optimal design methods for nonlinear models, still requires a solid prior knowledge of the reference curve parameter true values. Future research on this topic may apply Bayesian theory to this method, in order to account for the uncertainty associated with the prior estimates.

Finally, the similarity tests comparison presented in Chapter 6 is performed with a fixed sample size. Future considerations will assess the effect of adding replicates at each concentrations.

7.3 Conclusion

The introduction of the Hyper-Rectangle Test (HRT, see Section 1.2.5.2), and equivalence tests for parallelism in general, have significant advantages compared to p-value based tests to assess similarity between two log(response) curve. However, these methods are not without inherent challenges. Our proposed ellipsoid test, combined with our proposed zone of declared similarity derivation method, address some of the remaining challenges and move towards fit for purpose parallelism testing. Our recommendation, until the future work suggested in Section 7.2 is completed, is the use of the ellipsoid test when an experienced statistician is available to derive and implement the test acceptance limits. When it is not possible, the empirical $RSSE_{nonpar}$ and Maximum Departure tests can provide adequate performance in some specific conditions. We recommend using the ROUT method for outlier analysis. We also recommend estimating the log(concentration)-response curves

using a concentration range twice as wide as needed to observe 99% of the reference response range, with at least 10 concentration support points.

Listing of Relevant Publications

- E. Rozet, P. Lebrun, J.-F. Michiels, **P. Sondag**, T. Scherder, and B. Boulanger, "Analytical procedure validation and the quality by design paradigm," *Journal of biopharmaceutical statistics*, vol. 25, no. 2, pp. 260–268, 2015.

- **P. Sondag**, R. Joie, and H. Yang, "Comment and completion: implementation of parallelism testing for four-parameter logistic model in bioassays," *PDA journal of pharmaceutical science and technology*, vol. 69, no. 4, pp. 467–470, 2015.

- **P. Sondag**, P. Lebrun, E. Rozet, and B. Boulanger, "Assay validation," in *Nonclinical Statistics for Pharmaceutical and Biotechnology Industries*, pp. 415–432, Springer, 2016.

- S. Novick, **P. Sondag**, T. Schofield, and K. Miller, "A novel method for qualification of a potency assay through partial computer simulation," *PDA journal of pharmaceutical science and technology*, vol. 72, no. 3, pp. 249–263, 2018.

- M. Azadeh, B. Gorovits, J. Kamerud, S. MacMannis, A. Safavi, J. Sailstad, and **P. Sondag**, "Calibration curves in quantitative ligand binding assays: recommendations and best practices for preparation, design, and editing of calibration curves," *The AAPS journal*, vol. 20, no. 1, p. 22, 2018.

- **P. Sondag**, L. Zeng, B. Yu, R. Rousseau, B. Boulanger, H. Yang, and S. Novick, "Effect of a statistical outlier in potency bioassays," *Pharmaceutical statistics*, vol. 17, no. 6, pp. 701–709, 2018.

- M. Azadeh, **P. Sondag**, Y. Wang, M. Raines, and J. Sailstad, "Quality controls in ligand binding assays: Recommendations and best practices for preparation, qualification, maintenance of lot to lot consistency, and prevention of assay drift," *The AAPS journal*, vol. 21, no. 5, p. 89, 2019.

- **P. Sondag**, L. Zeng, B. Yu, H. Yang, and S. Novick, "Comparisons of outlier tests for potency bioassays," *Pharmaceutical statistics*, 2019.

- **P. Sondag** and P. Lebrun, "Risk-based similarity testing for potency assays using MCMC simulations," *Statistics in Biopharmaceutical Research*, 2020, *Manuscript accepted for publication, in-press.*

References

[1] A. Smith, *Oxford Dictionary of Biochemistry and Molecular Biology: Revised Edition*. Oxford University Press, 2000.

[2] R. R. Neubig, M. Spedding, T. Kenakin, and A. Christopoulos, "International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on terms and symbols in quantitative pharmacology," *Pharmacological reviews*, vol. 55, no. 4, pp. 597–606, 2003.

[3] G. E. Dinse, "An em algorithm for fitting a four-parameter logistic model to binary dose-response data," *Journal of agricultural, biological, and environmental statistics*, vol. 16, no. 2, pp. 221–232, 2011.

[4] D. Finney, "Radioligand assay," *Biometrics*, pp. 721–740, 1976.

[5] D. Lansky, "Strategic/modular bioassay design and analysis," USP's 7th Bioassay Workshop – Bioassay Life Cycle Approach, 2017.

[6] B. K. Lundholt, K. M. Scudder, and L. Pagliaro, "A simple technique for reducing edge effect in cell-based assays," *Journal of biomolecular screening*, vol. 8, no. 5, pp. 566–570, 2003.

[7] M. N. Khan and J. W. Findlay, *Ligand-binding assays: development, validation, and implementation in the drug development arena*. John Wiley & Sons, 2009.

[8] T. D. Pollard, "A guide to simple and informative binding assays," *Molecular biology of the cell*, vol. 21, no. 23, pp. 4061–4067, 2010.

[9] A. P. Davenport and F. D. Russell, "Radioligand binding assays: theory and practice," in *Current directions in radiopharmaceutical research and development*, pp. 169–179, Springer, 1996.

[10] S. Enna and M. Williams, "Challenges in the search for drugs to treat central nervous system disorders," *Journal of Pharmacology and Experimental Therapeutics*, vol. 329, no. 2, pp. 404–411, 2009.

[11] D. B. Bylund and S. Enna, "Receptor binding assays and drug discovery," in *Advances in Pharmacology*, vol. 82, pp. 21–34, Elsevier, 2018.

[12] W. Wang, "Potency testing of biopharmaceutical products." American Pharmaceutical Review, 2014.

[13] S. O. Doronina, B. E. Toki, M. Y. Torgov, B. A. Mendelsohn, C. G. Cerveny, D. F. Chace, R. L. DeBlanc, R. P. Gearing, T. D. Bovee, C. B. Siegall, *et al.*, "Development of potent monoclonal antibody auristatin conjugates for cancer therapy," *Nature biotechnology*, vol. 21, no. 7, pp. 778–784, 2003.

[14] E. L. Sievers and P. D. Senter, "Antibody-drug conjugates in cancer therapy," *Annual review of medicine*, vol. 64, pp. 15–29, 2013.

[15] H. Schluesener, R. Sobel, C. Linington, and H. Weiner, "A monoclonal antibody against a myelin oligodendrocyte glycoprotein induces relapses and demyelination in central nervous system autoimmune disease.," *The Journal of Immunology*, vol. 139, no. 12, pp. 4016–4021, 1987.

[16] L. D. Petz, "Cold antibody autoimmune hemolytic anemias," *Blood reviews*, vol. 22, no. 1, pp. 1–15, 2008.

[17] H. Rang, M. Dale, J. Ritter, and P. Moore, "Pharmacology (5th edn)," *Edinburgh: Churchill Livingstone*, 2003.

[18] M. R. Brunetto, F. Oliveri, G. Rocca, D. Criscuolo, E. Chiaberge, M. Capalbo, E. David, G. Verme, and F. Bonino, "Natural course and response to interferon of chronic hepatitis b accompanied by antibody to hepatitis b e antigen," *Hepatology*, vol. 10, no. 2, pp. 198–202, 1989.

[19] E. Tansey and P. Catterall, "Monoclonal antibodies: a witness seminar in contemporary medical history," *Medical history*, vol. 38, no. 3, pp. 322–327, 1994.

[20] X.-Y. Wang, B. Wang, and Y.-M. Wen, "From therapeutic antibodies to immune complex vaccines," *NPJ vaccines*, vol. 4, no. 1, p. 2, 2019.

[21] C. E. Hioe, M. L. Visciano, R. Kumar, J. Liu, E. A. Mack, R. E. Simon, D. N. Levy, and M. Tuen, "The use of immune complex vaccines to enhance antibody responses against neutralizing epitopes on hiv-1 envelope gp120," *Vaccine*, vol. 28, no. 2, pp. 352–360, 2009.

[22] B. Poirier, P. Variot, P. Delourme, J. Maurin, and S. Morgeaux, "Would an in vitro elisa test be a suitable alternative potency method to the in vivo immunogenicity assay commonly used in the context of international hepatitis a vaccines batch release?," *Vaccine*, vol. 28, no. 7, pp. 1796–1802, 2010.

[23] M. L. C. Cuervo, A. L. Sterling, I. A. Nicot, M. G. Rodríguez, and O. R. García, "Validation of a new alternative for determining in vitro potency in vaccines containing hepatitis b from two different manufacturers," *Biologicals*, vol. 36, no. 6, pp. 375–382, 2008.

[24] Y. Zhu, T. Zhang, J. Zhao, Z. Weng, Q. Yuan, S. Li, J. Zhang, N.-S. Xia, and Q. Zhao, "Toward the development of monoclonal antibody-based assays to probe virion-like epitopes in hepatitis b vaccine antigen," *Human vaccines & immunotherapeutics*, vol. 10, no. 4, pp. 1013–1023, 2014.

[25] M. Shank-Retzlaff, F. Wang, T. Morley, C. Anderson, M. Hamm, M. Brown, K. Rowland, G. Pancari, J. Zorman, R. Lowe, *et al.*, "Correlation between mouse potency and in vitro relative potency for human papillomavirus type 16 virus-like particles and gardasil® vaccine samples," *Human Vaccines*, vol. 1, no. 5, pp. 191–197, 2005.

[26] A. K. Pandey, R. K. Varshney, H. K. Sudini, and M. K. Pandey, "An improved enzyme-linked immunosorbent assay (elisa) based protocol using seeds for detection of five major peanut allergens ara h 1, ara h 2, ara h 3, ara h 6, and ara h 8," *Frontiers in nutrition*, vol. 6, 2019.

[27] X. Weng, G. Gaur, and S. Neethirajan, "Rapid detection of food allergens by microfluidics elisa-based optical sensor," *Biosensors*, vol. 6, no. 2, p. 24, 2016.

[28] R. Marlink, J. S. Allan, M. F. Mclane, M. Essex, K. C. Anderson, and J. E. Groopman, "Low sensitivity of elisa testing in early hiv infection," *New England Journal of Medicine*, vol. 315, no. 24, 1986.

[29] B. J. Johnson, K. E. Robbins, R. E. Bailey, B.-L. Cao, S. L. Sviat, R. B. Craven, L. W. Mayer, and D. T. Dennis, "Serodiagnosis of lyme disease: accuracy of a two-step approach using a flagella-based elisa and immunoblotting," *Journal of Infectious Diseases*, vol. 174, no. 2, pp. 346–353, 1996.

[30] M. Thouless, G. Beards, and T. Flewett, "Serotyping and subgrouping of rotavirus strains by the elisa test," *Archives of Virology*, vol. 73, no. 3-4, pp. 219–230, 1982.

[31] G. Kuno, I. Gomez, and D. Gubler, "An elisa procedure for the diagnosis of dengue infections," *Journal of virological methods*, vol. 33, no. 1-2, pp. 101–113, 1991.

[32] W. Duermeyer, F. Wielaard, and J. Van der Veen, "A new principle for the detection of specific igm antibodies applied in an elisa for hepatitis a," *Journal of Medical Virology*, vol. 4, no. 1, pp. 25–32, 1979.

[33] W. M. Hoskins and R. Craig, "Uses of bioassay in entomology," *Annual Review of Entomology*, vol. 7, no. 1, pp. 437–464, 1962.

[34] D. J. Finney, *Probit analysis: a statistical treatment of the sigmoid response curve*. Cambridge university press, Cambridge, 1952.

[35] S. Dudal, D. Baltrukonis, R. Crisino, M. J. Goyal, A. Joyce, K. Österlund, J. Smeraglia, Y. Taniguchi, and J. Yang, "Assay formats: recommendation for best practices and harmonization from the global bioanalysis consortium harmonization team," *The AAPS journal*, vol. 16, no. 2, pp. 194–205, 2014.

[36] A. Azizi, M. Tang, L. Gisonni-Lex, and L. Mallet, "Evaluation of infectious titer in a candidate hsv type 2 vaccine by a quantitative molecular approach," *BMC microbiology*, vol. 13, no. 1, p. 284, 2013.

[37] J. Schalk, C. de Vries, and P. Jongen, "Potency estimation of measles, mumps and rubella trivalent vaccines with quantitative pcr infectivity assay," *Biologicals*, vol. 33, no. 2, pp. 71–79, 2005.

[38] T. Ranheim, N. Mozier, and W. Egan, "Vaccine potency assays," in *Vaccine Analysis: Strategies, Principles, and Control*, pp. 521–541, Springer, 2015.

[39] European Parliament, "Directive 2010/63/eu." Official Journal of the European Union, 2010.

[40] C. F. Hendriksen, "Replacement, reduction and refinement alternatives to animal use in vaccine potency measurement," *Expert review of vaccines*, vol. 8, no. 3, pp. 313–322, 2009.

[41] G. Saha, "Design and analysis for bioassays," in *Design Workshop Lecture Notes, ISI, Kolkata*, pp. 61–76, 2002.

[42] Z. Govindarajulu, *Statistical techniques in bioassay*. Karger Medical and Scientific Publishers, 2001.

[43] K. Bailiss and T. Hill, "Biological assays for gibberellins," *The Botanical Review*, vol. 37, no. 4, pp. 437–479, 1971.

[44] H. Takigami, G. Suzuki, and S.-i. Sakai, "Application of bioassays for the detection of dioxins and dioxin-like compounds in wastes and the environment," *Interdisciplinary Studies on Environmental Chemistry: Biological Responses to Chemical Pollution (Murakami, Y., Nakayama, K., Kitamura, S.-I., Iwata, H., and Tanabe, S., Eds). Terrapub*, pp. 87–94, 2008.

[45] G. Suzuki, M. Nakamura, C. Michinaka, N. Tue, H. Handa, and H. Takigami, "Dioxin-like activity of brominated dioxins as individual compounds or mixtures in in vitro reporter gene assays with rat and mouse hepatoma cell lines," *Toxicology in Vitro*, vol. 44, pp. 134–141, 2017.

[46] M. Schäfer, S. Challand, E. Schick, S. Bader, D. Hainzl, K. Heinig, L. Müller, A. Papadimitriou, and J. Heinrich, "An ex vivo potency assay to assess active drug levels of a glp-1 agonistic peptide during

preclinical safety studies," *Bioanalysis*, vol. 7, no. 24, pp. 3063–3072, 2015.

[47] United States Pharmacopeial Convention, "<1032> design and development of biological assays," 2012.

[48] S. Jeffcoate, "The role of bioassays in the development, licensing and batch control of biotherapeutics," *Trends in biotechnology*, vol. 14, no. 4, pp. 121–124, 1996.

[49] J. Salvador, J. Adrian, R. Galve, D. Pinacho, M. Kreuzer, F. Sánchez-Baeza, and M. Marco, "Application of bioassays/biosensors for the analysis of pharmaceuticals in environmental samples," in *Comprehensive Analytical Chemistry* (M. Petrovic and D. Barcel, eds.), ch. 2.8, pp. 279–334, Elsevier, 2007.

[50] D. Rodbard and D. Hutt, "Statistical analysis of radioimmunoassays and immunoradiometric (labelled antibody) assays: a generalized weighted, iterative, least-square method for logistic curve fitting," in *Radioimmunoassay and related procedures in medicine*, 1974.

[51] D. A. Ratkowsky and T. J. Reedy, "Choosing near-linear parameters in the four-parameter logistic model for radioligand and related assays," *Biometrics*, pp. 575–582, 1986.

[52] G. E. Dinse and D. M. Umbach, "Quantifying relative potency in dose-response studies," in *Risk Assessment and Evaluation of Predictions*, pp. 315–331, Springer, 2013.

[53] European Pharmacopoeia, "Chapter 5.3. statistical analysis," 2004.

[54] G. Kärber, "Beitrag zur kollektiven behandlung pharmakologischer reihenversuche," *Naunyn-Schmiedebergs Archiv für experimentelle Pathologie und Pharmakologie*, vol. 162, pp. 480–483, Jul 1931.

[55] L. J. Reed and H. Muench, "A simple method of estimating fifty per cent endpoints," *American journal of epidemiology*, vol. 27, no. 3, pp. 493–497, 1938.

[56] B. W. Brown Jr, "Planning a quantal assay of potency," *Biometrics*, pp. 322–329, 1966.

[57] M. Azadeh, B. Gorovits, J. Kamerud, S. MacMannis, A. Safavi, J. Sailstad, and P. Sondag, "Calibration curves in quantitative ligand binding assays: recommendations and best practices for preparation, design, and editing of calibration curves," *The AAPS journal*, vol. 20, no. 1, p. 22, 2018.

[58] E. Bortolotto, R. Rousseau, B. Teodorescu, A. Wielant, and G. Debauve, "Assessing similarity with parallel-line and parallel-curve models: implementing the usp development/validation approach to a relative potency assay," *Bioprocess Int*, vol. 13, no. 6, 2015.

[59] G. D. Djira, "Relative potency estimation in parallel-line assays–method comparison and some extensions," *Communications in Statistics—Theory and Methods*, vol. 39, no. 7, pp. 1180–1189, 2010.

[60] N. Draper and H. Smith, *Applied regression analysis*. New York: Wiley, 1966.

[61] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.

[62] H. Benaroya, S. M. Han, and M. Nagurka, *Probability models in engineering and science*. CRC press, 2005.

[63] S. Clark and J. Perry, "Small sample estimation for taylor's power law," *Environmental and Ecological Statistics*, vol. 1, no. 4, pp. 287–302, 1994.

[64] J. G. Wagner, *Fundamentals of clinical pharmacokinetics*. Drug Intelligence Publications, 1975.

[65] "Jmp version 14.0." http://www.jmp.com/.

[66] "Softmax pro version 7." https://www.moleculardevices.com/.

[67] "Pla version 3.0." https://www.bioassay.de/.

[68] "Prism version 8." https://www.graphpad.com/.

[69] C. Ritz, F. Baty, J. C. Streibig, and D. Gerhard, "Dose-response analysis using r," *PloS one*, vol. 10, no. 12, p. e0146021, 2015.

[70] R. D. Wolfinger, "Fitting nonlinear mixed models with the new nlmixed procedure," in *Proceedings of the 24th Annual SAS Users Group International Conference (SUGI 24)*, pp. 278–284, 1999.

[71] R. I. Jennrich, "Asymptotic properties of non-linear least squares estimators," *The Annals of Mathematical Statistics*, vol. 40, no. 2, pp. 633–643, 1969.

[72] D. Katz, S. Azen, and A. Schumitzky, "Bayesian approach to the analysis of nonlinear models: implementation and evaluation," *Biometrics*, pp. 137–142, 1981.

[73] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," 1970.

[74] R. M. Neal *et al.*, "Mcmc using hamiltonian dynamics," *Handbook of markov chain monte carlo*, vol. 2, no. 11, p. 2, 2011.

[75] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of statistical software*, vol. 76, no. 1, 2017.

[76] M. Plummer *et al.*, "Jags: A program for analysis of bayesian graphical models using gibbs sampling," in *Proceedings of the 3rd international workshop on distributed statistical computing*, vol. 124, p. 10, Vienna, Austria., 2003.

[77] F. Chen, "Bayesian modeling using the mcmc procedure," in *Proceedings of the SAS Global Forum 2008 Conference, Cary NC: SAS Institute Inc*, 2009.

[78] D. Finney, "Statistical methods in biological assay. 2nd edn., section 21.5," 1964.

[79] P. G. Gottschalk and J. R. Dunn, "Measuring parallelism, linearity, and relative potency in bioassay and immunoassay data," *Journal of biopharmaceutical statistics*, vol. 15, no. 3, pp. 437–463, 2005.

[80] D. Lansky, "Strategic bioassay design, development, analysis, and validation," in *Statistics for Biotechnology Process Development*, pp. 131–156, Chapman and Hall/CRC, 2018.

[81] M. Davidian and D. M. Giltinan, "Nonlinear models for repeated measurement data: an overview and update," *Journal of agricultural, biological, and environmental statistics*, vol. 8, no. 4, p. 387, 2003.

[82] W. W. Hauck, R. C. Capen, J. D. Callahan, J. E. De Muth, H. Hsu, D. Lansky, N. C. Sajjadi, S. S. Seaver, R. R. Singer, and D. Weisman, "Assessing parallelism prior to determining relative potency," *PDA Journal of Pharmaceutical Science and Technology*, vol. 59, no. 2, pp. 127–137, 2005.

[83] J. N. Jonkman and K. Sidik, "Equivalence testing for parallelism in the four-parameter logistic model," *Journal of biopharmaceutical statistics*, vol. 19, no. 5, pp. 818–837, 2009.

[84] H. Yang, H. J. Kim, L. Zhang, R. J. Strouse, M. Schenerman, and X.-R. Jiang, "Implementation of parallelism testing for four-parameter logistic model in bioassays," *PDA journal of pharmaceutical science and technology*, vol. 66, no. 3, pp. 262–269, 2012.

[85] P. Sondag, R. Joie, and H. Yang, "Comment and completion: implementation of parallelism testing for four-parameter logistic model in bioassays," *PDA journal of pharmaceutical science and technology*, vol. 69, no. 4, pp. 467–470, 2015.

[86] R. L. Berger and J. C. Hsu, "Bioequivalence trials, intersection-union tests and equivalence confidence sets," *Statistical Science*, vol. 11, no. 4, pp. 283–319, 1996.

[87] D. Lansky, "Near-universal equivalence bounds for similarity in bioassays," ASA Biopharmaceutical Section Nonclinical Biostatistics Conference, 2019.

[88] S. J. Novick, H. Yang, and J. J. Peterson, "A bayesian approach to parallelism testing in bioassay," *Statistics in biopharmaceutical research*, vol. 4, no. 4, pp. 357–374, 2012.

[89] S. J. Novick, X. Zhang, and H. Yang, "A new pk equivalence test for a bridging study," *Journal of biopharmaceutical statistics*, vol. 26, no. 5, pp. 992–1002, 2016.

[90] S. Novick and H. Yang, "A fast and reliable test for parallelism in bioassay," *Journal of biopharmaceutical statistics*, pp. 1–13, 2019.

[91] E. C. Fieller, "A fundamental formula in the statistics of biological assay, and some applications," *Quart. J. Pharm*, vol. 17, pp. 117–123, 1944.

[92] E. C. Fieller, "Some problems in interval estimation," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 16, no. 2, pp. 175–185, 1954.

[93] International Conference on Harmonisation, "Q8: Pharmaceutical development," 2009.

[94] P. Borman, P. Nethercote, M. Chatfield, D. Thompson, and K. Truman, "The application of quality by design to analytical methods," 2007.

[95] M. Schweitzer, M. Pohl, M. Hanna-Brown, P. Nethercote, P. Borman, G. Hansen, K. Smith, and J. Larew, "Implications and opportunities of applying qbd principles to analytical measurements," *Pharmaceutical Technology*, vol. 34, no. 2, pp. 52–59, 2010.

[96] P. Nethercote and J. Ermer, "Quality by design for analytical methods: implications for method validation and transfer," *Pharmaceutical Technology*, vol. 36, no. 10, pp. 74–79, 2012.

[97] E. Rozet, P. Lebrun, J.-F. Michiels, P. Sondag, T. Scherder, and B. Boulanger, "Analytical procedure validation and the quality by design paradigm," *Journal of biopharmaceutical statistics*, vol. 25, no. 2, pp. 260–268, 2015.

[98] P. Sondag, P. Lebrun, E. Rozet, and B. Boulanger, "Assay validation," in *Nonclinical Statistics for Pharmaceutical and Biotechnology Industries*, pp. 415–432, Springer, 2016.

[99] H. Yang and L. Zhang, "Evaluations of parallelism testing methods using roc analysis," *Statistics in Biopharmaceutical Research*, vol. 4, no. 2, pp. 162–173, 2012.

[100] V. Chew, "Confidence, prediction, and tolerance regions for the multivariate normal distribution," *Journal of the American Statistical Association*, vol. 61, no. 315, pp. 605–617, 1966.

[101] C. E. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita'.," 1936.

[102] J. Kruschke, *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[103] F. Feng, A. P. Sales, and T. B. Kepler, "A bayesian approach for estimating calibration curves and unknown concentrations in immunoassays," *Bioinformatics*, vol. 27, no. 5, pp. 707–712, 2011.

[104] K. Klauenberg, B. Ebert, J. Voigt, M. Walzel, J. E. Noble, A. E. Knight, and C. Elster, "Bayesian analysis of an international elisa comparability study," *Clinical chemistry and laboratory medicine*, vol. 49, no. 9, pp. 1459–1468, 2011.

[105] K. Klauenberg, M. Walzel, B. Ebert, and C. Elster, "Informative prior distributions for elisa analyses," *Biostatistics*, vol. 16, no. 3, pp. 454–464, 2015.

[106] K. Sidik and J. N. Jonkman, "Testing for parallelism in the heteroscedastic four-parameter logistic model," *Journal of biopharmaceutical statistics*, vol. 26, no. 2, pp. 250–268, 2016.

[107] P. Lebrun, P. Sondag, X. Lories, J.-F. Michiels, E. Rozet, and B. Boulanger, "Quality by design applied in formulation development and robustness," in *Statistics for Biotechnology Process Development*, pp. 77–92, Chapman and Hall/CRC, 2018.

[108] International Conference on Harmonisation, "Q9: Quality risk management," 2006.

[109] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[110] K. V. Mardia, "Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies," *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 115–128, 1974.

[111] Y. Zhao, C. Genest, *et al.*, "Inference for elliptical copula multivariate response regression models," *Electronic Journal of Statistics*, vol. 13, no. 1, pp. 911–984, 2019.

[112] A. Sklar, A. SKLAR, and C. Sklar, "Fonctions de repartition à n dimensions et leurs marges," 1959.

[113] P. Sondag, L. Zeng, B. Yu, R. Rousseau, B. Boulanger, H. Yang, and S. Novick, "Effect of a statistical outlier in potency bioassays," *Pharmaceutical statistics*, vol. 17, no. 6, pp. 701–709, 2018.

[114] International Conference on Harmonisation, "Q6: Specifications," 1999.

[115] A. VØlund, "Application of the four-parameter logistic model to bioassay: comparison with slope ratio and parallel line models," *Biometrics*, pp. 357–365, 1978.

[116] V. Barnett and T. Lewis, *Outliers in statistical data*. Wiley, 1974.

[117] K. Fleetwood, "Accounting for variability in relative potency estimates," BEBPA's 5th Annual Biological Assay Conference, 2012.

[118] T. Schofield, "Lifecycle approach to bioassay," in *Nonclinical statistics for pharmaceutical and biotechnology industries*, pp. 433–460, Springer, 2016.

[119] R. B. Dean and W. Dixon, "Simplified statistics for small numbers of observations," *Analytical Chemistry*, vol. 23, no. 4, pp. 636–638, 1951.

[120] H. J. Motulsky and R. E. Brown, "Detecting outliers when fitting data with nonlinear regression–a new method based on robust nonlinear regression and the false discovery rate," *BMC bioinformatics*, vol. 7, no. 1, p. 123, 2006.

[121] P. Sondag, L. Zeng, B. Yu, H. Yang, and S. Novick, "Comparisons of outlier tests for potency bioassays," *Pharmaceutical statistics*, 2019.

[122] United States Pharmacopeial Convention, "<1010> analytical data interpretation and treatment," 2012.

[123] F. R. Hampel, "The breakdown points of the mean combined with some rejection rules," *Technometrics*, vol. 27, no. 2, pp. 95–107, 1985.

[124] B. Rosner, "On the detection of many outliers," *Technometrics*, vol. 17, no. 2, pp. 221–227, 1975.

[125] B. Rosner, "Percentage points for a generalized esd many-outlier procedure," *Technometrics*, vol. 25, no. 2, pp. 165–172, 1983.

[126] W. Dixon, "Processing data for outliers," *Biometrics*, vol. 9, no. 1, pp. 74–89, 1953.

[127] P. Huber, *Robust Statistics*. Wiley, 1981.

[128] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[129] S. Psarakis and J. Panaretoes, "The folded t distribution," *Communications in Statistics-Theory and Methods*, vol. 19, no. 7, pp. 2717–2734, 1990.

[130] A. Atkinson, "Developments in the design of experiments, correspondent paper," *International Statistical Review/Revue Internationale de Statistique*, pp. 161–177, 1982.

[131] A. Atkinson and A. Donev, "Optimum experimental designs. clarendon," 1992.

[132] J. Morgan and J. Stallings, "On the a criterion of experimental design," *Journal of Statistical Theory and Practice*, vol. 8, no. 3, pp. 418–422, 2014.

[133] J. P. Buonaccorsi, "Designs for slope ratio assays," *Biometrics*, pp. 875–882, 1986.

[134] J. Buonaccorsi and H. Iyer, "Optimal designs for ratios of linear combinations in the general linear model," *Journal of Statistical Planning and Inference*, vol. 13, pp. 345–356, 1986.

[135] A. R. G. Mukkula and R. Paulen, "Model-based design of optimal experiments for nonlinear systems in the context of guaranteed parameter estimation," *Computers & Chemical Engineering*, vol. 99, pp. 198–213, 2017.

[136] H. Chernoff, "Locally optimal designs for estimating parameters," *The Annals of Mathematical Statistics*, pp. 586–602, 1953.

[137] V. Melas, "On a functional approach to locally optimal designs," in *mODa 7—Advances in Model-Oriented Design and Analysis*, pp. 97–105, Springer, 2004.

[138] L. Kalish and J. Rosenberger, "Optimal designs for the estimation of the logistic function," in *Technical Report 33*, The Pennsylvania State University, Dept. of Statistics, State College PA, 1978.

[139] M. Bezeau and L. Endrenyi, "Design of experiments for the precise estimation of dose-response parameters: the hill equation," *Journal of theoretical biology*, vol. 123, no. 4, pp. 415–430, 1986.

[140] E. Masoudi, H. Holling, and W. K. Wong, "Application of imperialist competitive algorithm to find minimax and standardized maximin optimal designs," *Computational statistics & data analysis*, vol. 113, pp. 330–345, 2017.

[141] A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, "A fast and scalable method for a-optimal design of experiments for infinite-dimensional bayesian nonlinear inverse problems," *SIAM Journal on Scientific Computing*, vol. 38, no. 1, pp. A243–A272, 2016.

[142] L. A. Khinkis, L. Levasseur, H. Faessel, and W. R. Greco, "Optimal design for estimating parameters of the 4-parameter hill model," *Nonlinearity in biology, toxicology, medicine*, vol. 1, no. 3, p. 15401420390249925, 2003.

[143] N. François, B. Govaerts, and B. Boulanger, "Optimal designs for inverse prediction in univariate nonlinear calibration models," *Chemometrics and intelligent laboratory systems*, vol. 74, no. 2, pp. 283–292, 2004.

[144] M. D. McKay, R. J. Beckman, and W. J. Conover, "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979.

[145] I. Burghaus and H. Dette, "Optimal designs for nonlinear regression models with respect to non-informative priors," *Journal of Statistical Planning and Inference*, vol. 154, pp. 12–25, 2014.

[146] P. Tusto, T. E. O'Brien, and M. Tiensuwan, "Optimal design strategies for relative potency using the two-parameter log-logistic model," *Model Assisted Statistics and Applications*, vol. 11, no. 2, pp. 109–123, 2016.

[147] S. Deming, *The 4PL*. Statistical Designs, 2015.

[148] A. DeLean, P. Munson, and D. Rodbard, "Simultaneous analysis of families of sigmoidal curves: application to bioassay, radioligand assay, and physiological dose-response curves.," *American Journal of Physiology-Endocrinology And Metabolism*, vol. 235, no. 2, p. E97, 1978.

[149] N. R. Draper and H. Smith, *Applied regression analysis*, vol. 326. John Wiley and Sons, 1998.

[150] K. Fleetwood, F. Bursa, and A. Yellowlees, "Parallelism in practice: approaches to parallelism in bioassays," *PDA journal of pharmaceutical science and technology*, vol. 69, no. 2, pp. 248–263, 2015.

[151] P. Sondag, "Equivalence margins to assess parallelism between 4pl curve," NonClinical Statistics Conference, 2014.

[152] S. Novick, P. Sondag, T. Schofield, and K. Miller, "A novel method for qualification of a potency assay through partial computer simulation," *PDA journal of pharmaceutical science and technology*, vol. 72, no. 3, pp. 249–263, 2018.

[153] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, "Bayesian data analysis chapman & hall," *CRC Texts in Statistical Science*, 2004.

[154] P. Faya, P. Sondag, S. Novick, D. Banton, J. Seaman, J. Stamey, and B. Boulanger, "The current state of bayesian methods in nonclinical pharmaceutical statistics: survey results and recommendations from the dia/asa-biop nonclinical bayesian working group," *Manuscript sumbitted for publication.*

[155] United States Pharmacopeial Convention, "<111> design and analysis of biological assays," 2001.

[156] US Food and Drug Administration, "Cfr-code of federal regulations title 21," 2017.

[157] M. Febrero, P. Galeano, and W. González-Manteiga, "Outlier detection in functional data by depth measures, with application to identify abnormal nox levels," *Environmetrics: The official journal of the International Environmetrics Society*, vol. 19, no. 4, pp. 331–345, 2008.

[158] A. Arribas-Gil and J. Romo, "Shape outlier detection and visualization for functional data: the outliergram," *Biostatistics*, vol. 15, no. 4, pp. 603–619, 2014.

[159] SAS Institute Inc., "The NLMIXED Procedure," in *SAS/STAT®15.1 User's Guide*, pp. 77–92, SAS Institute Inc., 2018.

[160] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[161] A. Gelman, "Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)," *Bayesian analysis*, vol. 1, no. 3, pp. 515–534, 2006.

[162] J. O. Berger, D. R. Insua, and F. Ruggeri, "Bayesian robustness," in *Robust Bayesian Analysis*, pp. 1–32, Springer, 2000.

[163] M. Lavine, "Sensitivity in bayesian statistics: the prior and the likelihood," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 396–399, 1991.

[164] S. Morita, P. F. Thall, and P. Müller, "Determining the effective sample size of a parametric prior," *Biometrics*, vol. 64, no. 2, pp. 595–602, 2008.

[165] Y. Gu, H.-L. Wei, and M. M. Balikhin, "Nonlinear predictive model selection and model averaging using information criteria," *Systems Science & Control Engineering*, vol. 6, no. 1, pp. 319–328, 2018.

This page is intentionally left blank

Appendices

This page is intentionally left blank

Appendix A

Bayesian 4PL Example

This Appendix presents an example of a 4PL curve fit, using historical data as prior information for a pre-validation study of a serial dilution assay. All data presented in this Appendix are simulated.

Consider the true values:

$$\begin{aligned} yMin &= 0 \\ yMax &= 1 \\ c &= 0.0625 \\ S &= 2 \\ \sigma_y^2 &= 0.0016 \\ \sigma_c^2 &= 0.1644 \\ \sigma_{yMax}^2 &= 0.01 \end{aligned}$$

Where $c = \log(C_{50})$, σ_c^2 and σ_{yMax}^2 are respectively the plate to plate variabilities affecting c and $yMax$, and σ_y^2 is the measurement variability.

A.1 Development Study

A development set of four plates is available as historical data. Each plate contains one serial dilution curve, obtained from three replicates at ten concentration points: $1, 0.5, 0.25, \ldots, 0.002$. The development data are presented

in Figure A.1, and the R code used to generate the data is available in Section A.5.

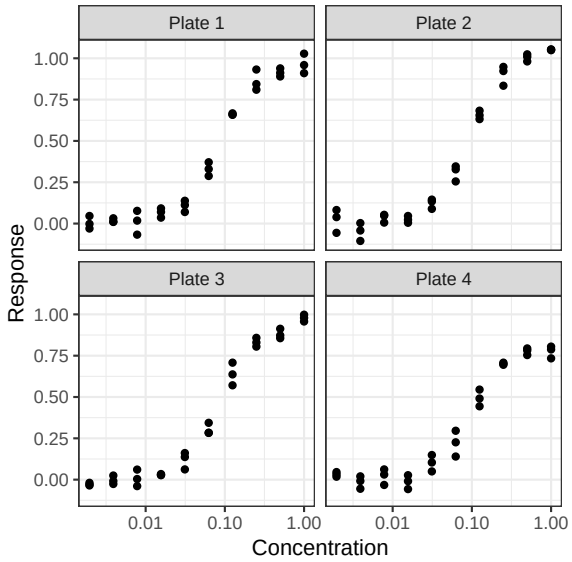


Figure A.1: Development set of plates

A 4PL model (Equation 1.2) is fit on the data via maximum likelihood, with a random effect of the plate affecting both the upper asymptote and c.

Table A.1 presents the estimates and approximated standard errors of the curve parameters. Table A.2 presents the estimates and approximated degrees of freedoms of the variance components. All estimates, the curve parameters standard errors, and the degrees of freedom of the residual variability were calculated using the R function `nlmer` from the package `lme4` [160]. The degrees of freedom of the between-plates variabilities were approximated by the number of available plates -1 [159].

Table A.1: Maximum likelihood estimates and standard error of curve parameters

Parameter	Estimate	Std. Error
$yMin$	0.001	0.006
$yMax$	0.942	0.047
c	-2.381	0.031
S	1.936	0.084

Table A.2: Maximum likelihood estimates and degrees of freedom of variance components

Variance Component	Estimate	Degrees of Freedom
σ^2_{yMax}	0.0083	3
σ^2_c	0.0013	3
σ^2_y	0.0016	113

A.2 Pre-validation study

A pre-validation set of nine plates is available. Each plate contains one serial dilution curve, obtained from three replicates at ten concentration points. The concentration support points were chosen using the methodology presented in Chapter 5, and the curve parameter estimates calculated in Section A.1. The maximum concentration and dilution factor were respectively chosen to be 6.7 and 3. The pre-validation data are presented in Figure A.2, and the R code used to generate the data is available in Section A.5.

As the goal is to be able to make predictions for the validation study, Bayesian methodologies are used. The model was fit with both informative and non-informative prior distributions for the curve parameters and variance components to assess the effect of leveraging prior knowledge.

For the curve parameters, Normal prior distributions were used. For variance components, inverse-Gamma distributions were used. Gelman (2006) advises against using the non-informative inverse-Gamma distributions [161]. However, they are a convenient choice in order to use the same code at different phases of a study, and simply update the prior information. If the inverse-Gamma choice is made, like in this case, visually evaluating the ob-

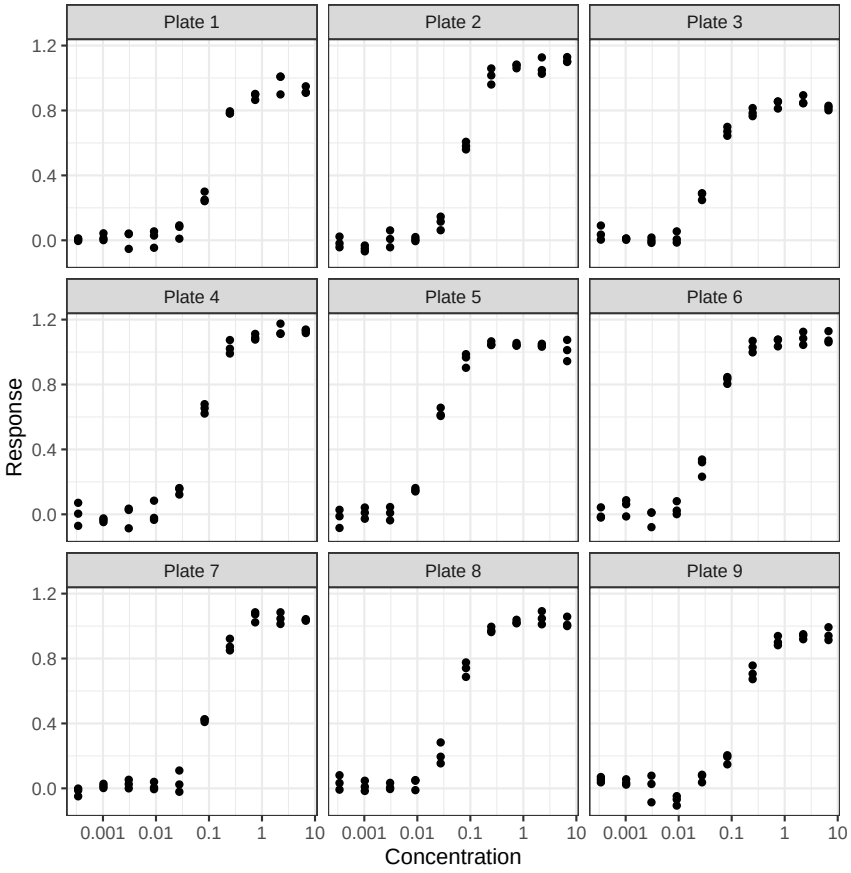


Figure A.2: Pre-validation set of plates

tained posterior distribution is particularly important.

For the non-informative Normal prior distributions, all means were set at 0 and all standard deviations were set at 1e5. For the informative Normal prior distributions, the means and standard deviations were set as the curve parameter estimates and approximated standard errors from the development study, respectively.

For the non-informative inverse-Gamma distributions, shapes and scales were set all at 1e-5. For the informative inverse-Gamma distributions, shapes were set as $0.5 \times df$ and scales were set as $0.5 \times df \times \hat{\sigma}$, for each variance component, where $\hat{\sigma}$ and df are respectively the maximum likelihood estimate

and approximated degrees of freedom from the development study.

In both cases, the joint posterior distribution of the model parameters was sampled by Hybrid MCMC using Stan. Four independent chains were run with each 30,000 draws, including a warm-up of 5,000 draws, thinning every 5 draws. This resulted in 5,000 posterior draws per chain and 30,000 posterior draws total.

Figures A.3 and A.4 respectively present the trace plots and density plots of each parameter marginal posterior distribution obtained with non-informative distributions. The estimated medians for the variance components marginal posterior distributions were 0.0088, 0.3991, and 0.0014 respectively for σ^2_{yMax}, σ^2_c, and σ^2_y. While σ^2_c seems over-estimated compared to the true value, the sample variance of the generated cs was actually 0.3333, so we consider the inverse-Gamma choice for non-informative prior to be reasonable.

Figures A.5 and A.7 respectively present the trace plots and density plots of each parameter marginal posterior distribution obtained with informative distributions. As expected, these posterior distributions were visibly narrower than when non-informative prior distributions were used.

At 1000 different concentration points across the pre-validation study concentration range, we then drew 30,000 samples from the posterior predictive distribution of the reference curve. Figure A.7 presents the estimated median and 95% highest posterior density interval (HPDI) of the posterior predictive distribution calculated using both non-informative and informative prior. Using informative prior distributions allowed to narrow the HPDI in the asymptotes, but most importantly in the middle of the curve, which is the area that is used to calculate the potency.

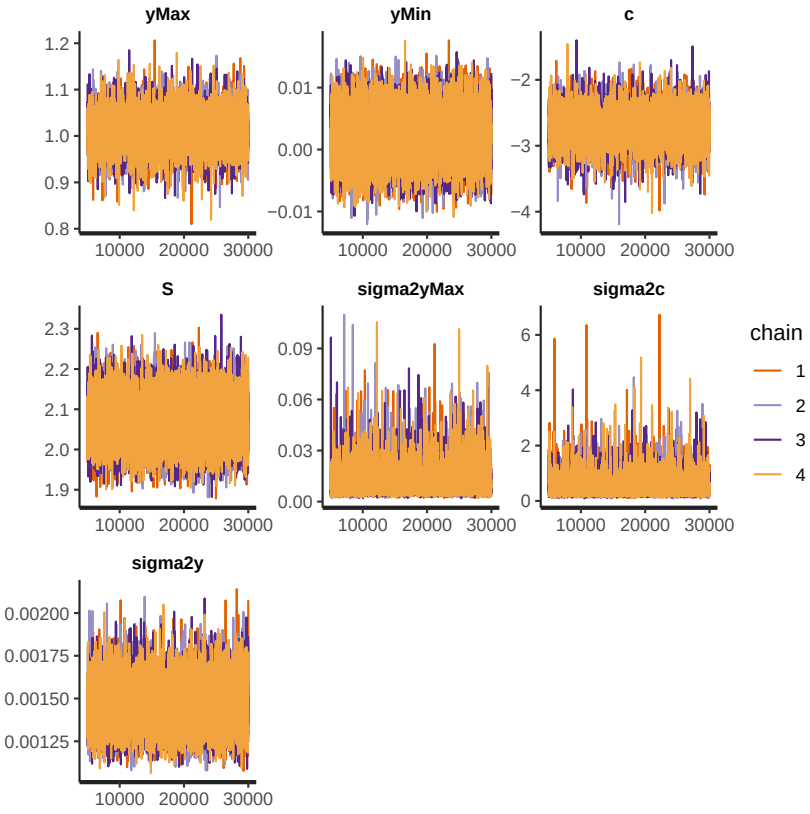


Figure A.3: Trace plot of each model parameter when using non-informative prior distributions.

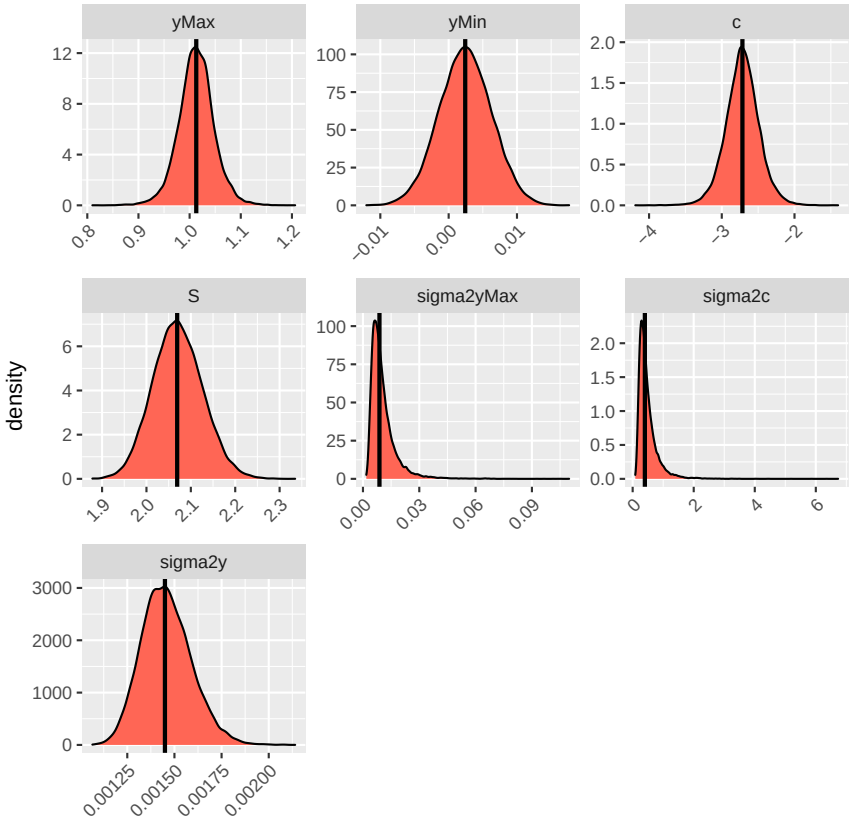


Figure A.4: Density plot of each model parameter when using non-informative prior distributions. The vertical lines are the estimated posterior medians.

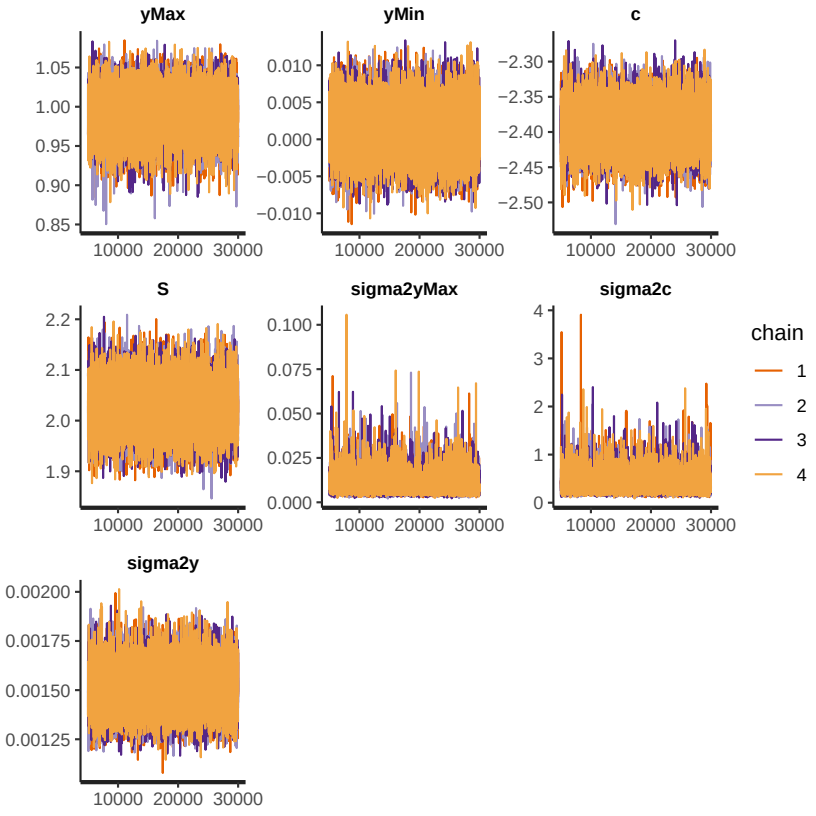


Figure A.5: Trace plot of each model parameter when using informative prior distributions.

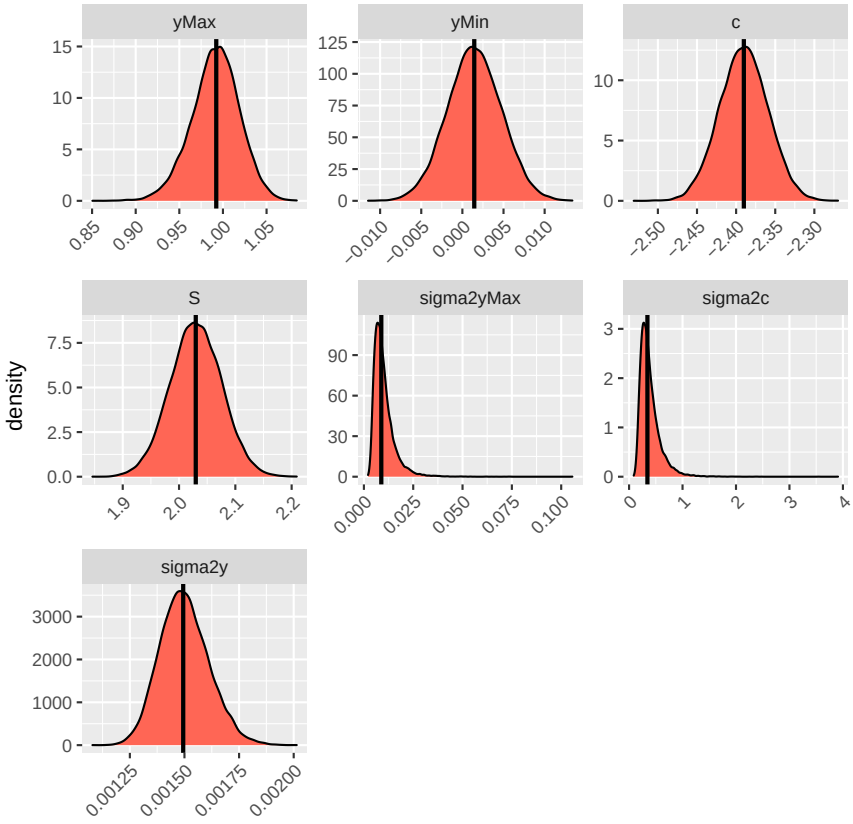


Figure A.6: Density plot of each model parameter when using informative prior distributions. The vertical lines are the estimated posterior medians.

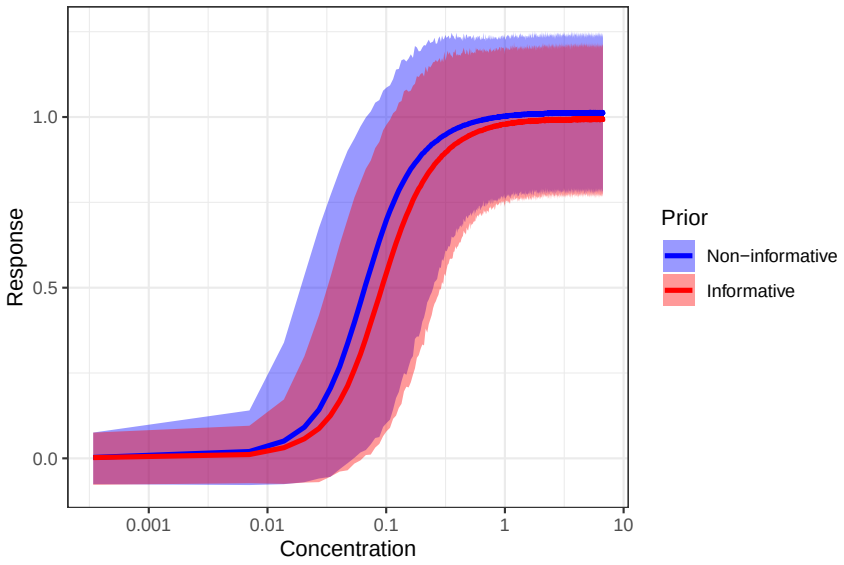


Figure A.7: Posterior predictive distribution of the whole serial dilution profile using non-informative and informative distributions. The plain lines represent the medians of the distribution, while the ribbons represent the 95% *HPDI.*

A.3 Discussion

Using informative prior distribution allows to narrow the uncertainty around the curve parameter estimates and future predictions. Chapter 2 also shows the positive effect of informative prior distributions on the derivation of acceptance criteria for similarity testing.

Note that, in this particular case, the development and pre-validation data were simulated from the exact same distributions. This may not be the case in practice, as study effects can occur, as well as changes in the assay or in the material. In case some changes happened, informative prior distributions could be used for only a subset of the curve parameters or variance components. Alternatively, power prior can be used to reduce the effect of the prior distribution on the posterior [164]. Prior sensitivity analysis should also be performed to find the most appropriate prior for the data [162, 163].

Another aspect is model selection. While formal methods exist to compare models [165], data visualization from an experienced statistician may be enough to decide what curve parameter is affected by a variance component.

A.4 Data Generation

The R code below may be used to generate the development study data:

```
set.seed(20200516)
## 4PL equation:
hillmod =
  function(
    x, # concentration
    yMin, # lower asymptote
    yMax, # upper asymptote
    lc50, # log(c50)
    S # steepness
  ){
    yMax +
```

```
      (yMin - yMax) /
      (1 + exp(S * (log(x) - lc50)))
  }


## True parameters

yMin = 0 # lower asymptote
yMax = 1 # upper asymptote
lc50 = log(0.0625) # log(c50)
S = 2 # steepness
sigma = 0.04 # residual SD
sigma_lc50 = log(1.5) # plate to plate log(c50) SD
sigma_ymax = 0.1 # plate to plate yMax SD

## Generate Dev Data

nplate_dev = 4 # number of development plates

# Random Effect: log(c50)
lc50_plate_dev =
  rnorm(
    n = nplate_dev,
    mean = lc50,
    sd = sigma_lc50
  )

# Random Effect: yMax
ymax_plate_dev =
  rnorm(
    n = nplate_dev,
    mean = yMax,
    sd = sigma_ymax
  )

# create data set
d_dev = expand.grid(
  plate = factor(1:nplate_dev),
```

```
  concentration = 1/(2^(0:9)),
  rep = 1:3
)
d_dev$y =
  round(
    rnorm(
      nrow(d_dev),
      hillmod(
        d_dev$concentration,
        yMin = yMin,
        yMax = ymax_plate_dev[d_dev$plate],
        lc50 = (lc50_plate_dev[d_dev$plate]),
        S = S
      ),
      sigma
    ),
    3
  )
```

The R code below may be used to generate the pre-validation study data:

```
set.seed(20200519)

nplate_qual = 9 # number of development plates

# Random Effect: log(c50)
lc50_plate_qual =
  rnorm(
    n = nplate_qual,
    mean = lc50,
    sd = sigma_lc50
  )

# Random Effect: yMax
ymax_plate_qual =
  rnorm(
```

```
    n = nplate_qual,
    mean = yMax,
    sd = sigma_ymax
  )

# create data set
d_qual = expand.grid(
  plate = factor(1:nplate_qual),
  concentration = 6.7/3^(0:9),
  rep = 1:3
)
d_qual$y =
  round(
    rnorm(
      nrow(d_qual),
      hillmod(
        d_qual$concentration,
        yMin = yMin,
        yMax = ymax_plate_qual[d_qual$plate],
        lc50 = (lc50_plate_qual[d_qual$plate]),
        S = S
      ),
      sigma
    ),
    3
  )
```

A.5 Stan Code

The Stan code below was used to fit the 4PL model on the pre-validation study data:

```
data {
  // number of observations
  int N;
```

```
  // numer of groups
  int n_groups;
  // vector of responses
  real y[N];
  // vector of concentrations
  real x[N];
  // vector of group
  int group[N];
  // prior mean for parameters in order yMin, yMax, c, S
  real mean_prior[4];
  // prior SD for parameters in order yMin, yMax, c, S
  real sd_prior[4];
  // prior alpha for VCs in order yMax, c, residuals
  real alpha_prior[3];
  // prior alpha for VCs in order yMax, c, residuals
  real beta_prior[3];
}
parameters {
  // lower asymptote
  real yMin;
  // upper asymptote
  real yMax;
  // Steepness
  real<lower=0> S;
  // log(c50)
  real c;
  // between group variance yMax
  real<lower=0> sigma2yMax;
  // random effect yMax
  real r_yMax[n_groups];
  // between group variance log(c50)
  real<lower=0> sigma2c;
  // random effect log(c50)
  real r_c[n_groups];
  // measurement variance
  real<lower=0> sigma2y;
}
model {
```

```
  // yMax by group
  real yMax_group[N];
  // log(c50) by group
  real c50_group[N];
  // predicted value
  real mu[N];

  // prior

  yMin ~ normal(mean_prior[1], sd_prior[1]);
  yMax ~ normal(mean_prior[2], sd_prior[2]);
  c ~ normal(mean_prior[3], sd_prior[3]);
  S ~ normal(mean_prior[4], sd_prior[4]);

  sigma2yMax ~ inv_gamma(alpha_prior[1], beta_prior[1]);
  sigma2c ~ inv_gamma(alpha_prior[2], beta_prior[2]);
  sigma2y ~ inv_gamma(alpha_prior[3], beta_prior[3]);


  // Compute random effects
  for(j in 1:n_groups){
    r_yMax[j]~normal(0, sqrt(sigma2yMax));
    r_c[j]~normal(0, sqrt(sigma2c));
  }

  // Fit model
  for(i in 1:N){
    yMax_group[i] = yMax + r_yMax[group[i]];
    c50_group[i] = c + r_c[group[i]];
    mu[i] = yMax_group[i] + (yMin - yMax_group[i])/
      (1 + exp(S * (log(x[i]) - c50_group[i])));
  }
  y ~ normal(mu, sqrt(sigma2y));
}

}
```

Appendix B

Effect of a Statistical Outlier when RP = 50% and 200%

B.1 Results for RP = 50%

This Appendix presents the supplementary material of Chapter 3. In the chapter, we only present results when $RP = 100\%$. This Appendix shows the same results for low and high RP. Figures B.1 to B.6 present the results when the C_{50} of the test product is double of the C_{50} of the reference product (true RP = 50%). In general, the results are similar to the ones found in Chapter 4, with a lateral shift that accounts for the different RP.

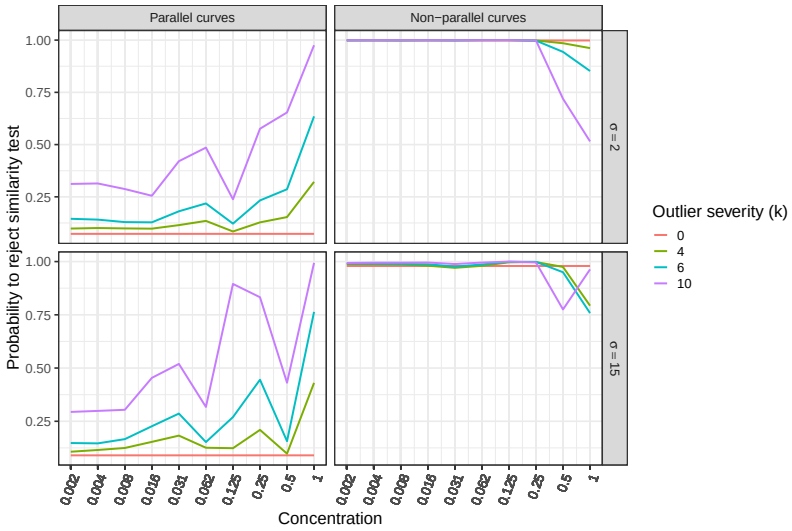


Figure B.1: Single observation outlier: effect on similarity testing when true RP = 50%

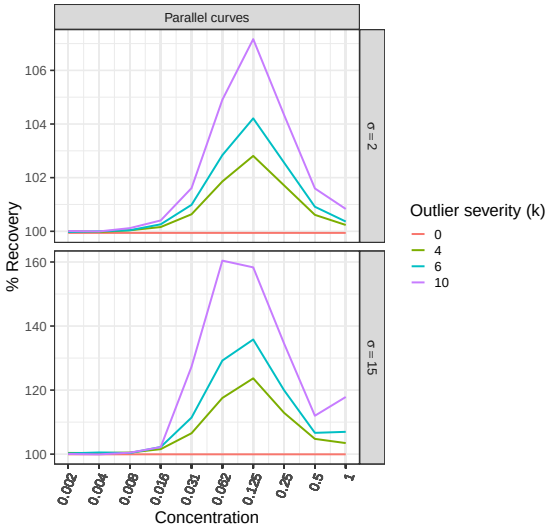


Figure B.2: Single observation outlier: geometric mean of the recovery when true RP = 50%

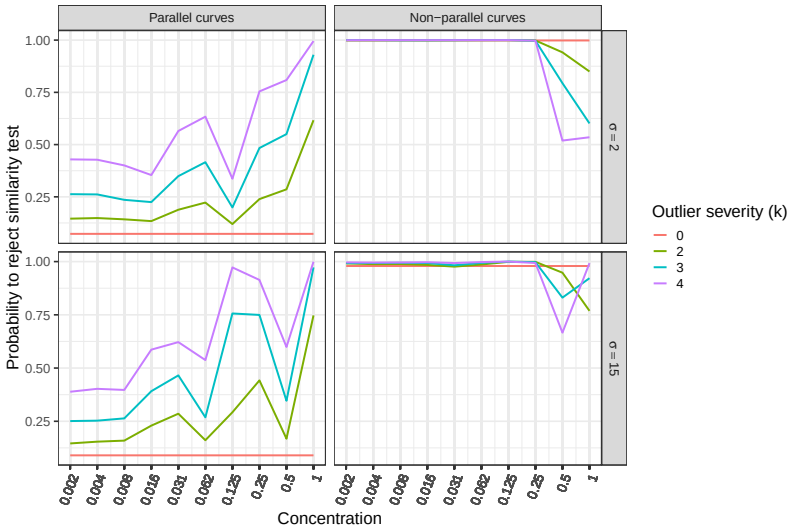


Figure B.3: Concentration point outlier: effect on similarity testing when true RP = 50%

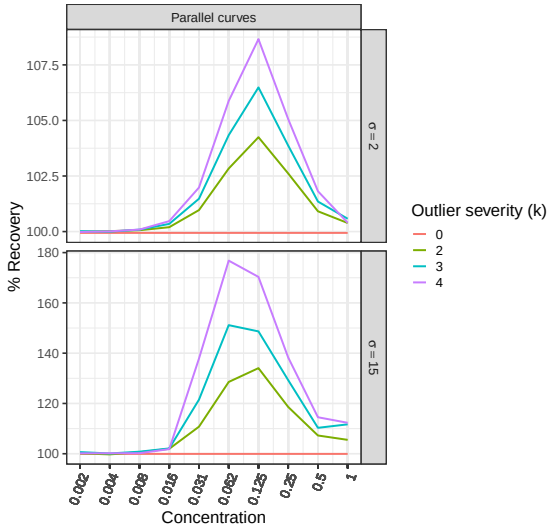


Figure B.4: Concentration point outlier: geometric mean of the recovery when true RP = 50%

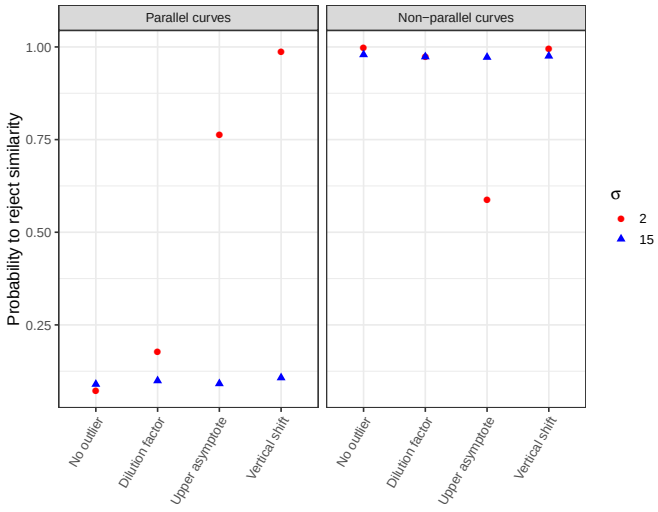


Figure B.5: Whole curve outlier: effect on similarity testing when true RP = 50%

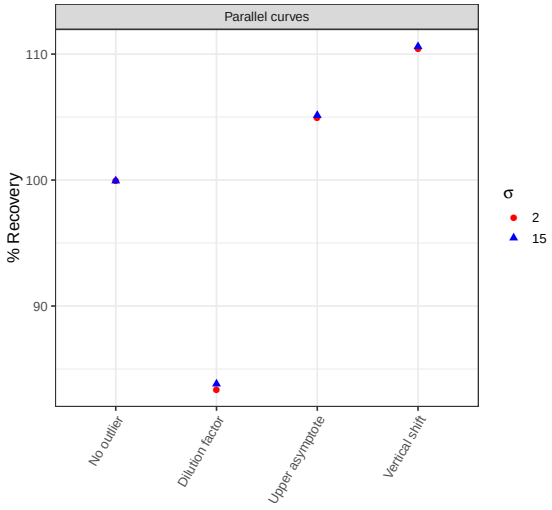


Figure B.6: Whole curve outlier: geometric mean of the recovery when true RP = 50%

B.2 Results for RP = 200%

Figures B.7 to B.12 present the results when the C_{50} of the test product is double of the EC50 of the reference product (true RP = 200%). In general, the results are similar to the ones found in Chapter 4, with a lateral shift that accounts for the different RP.

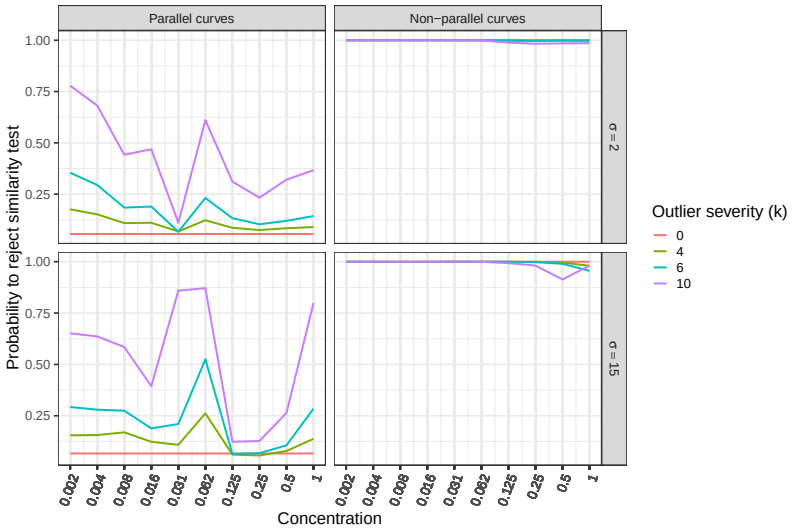


Figure B.7: Single observation outlier: effect on similarity testing when true RP = 200%

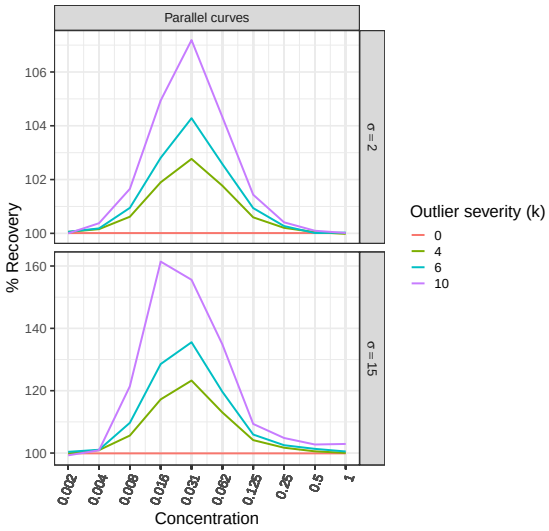


Figure B.8: Single observation outlier: geometric mean of the recovery when true RP = 200%

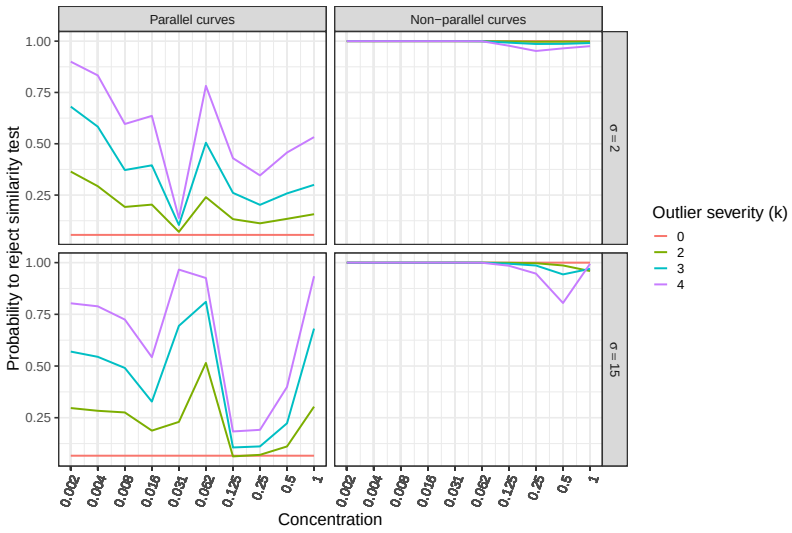


Figure B.9: Concentration point outlier: effect on similarity testing when true RP = 200%

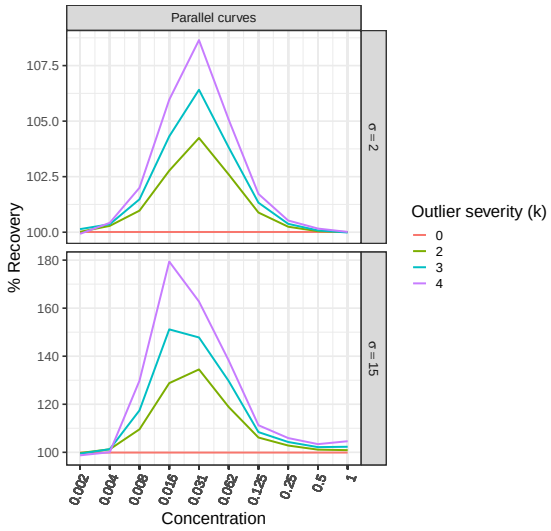


Figure B.10: Concentration point outlier: geometric mean of the recovery when true RP = 200%

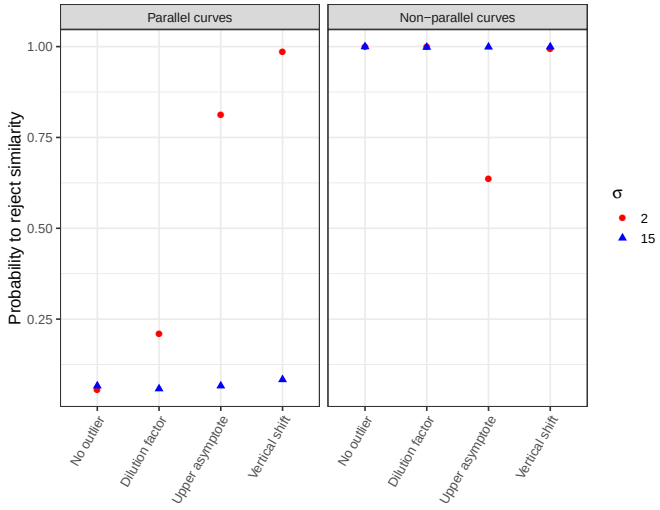


Figure B.11: Whole curve outlier: effect on similarity testing when true RP = 200%

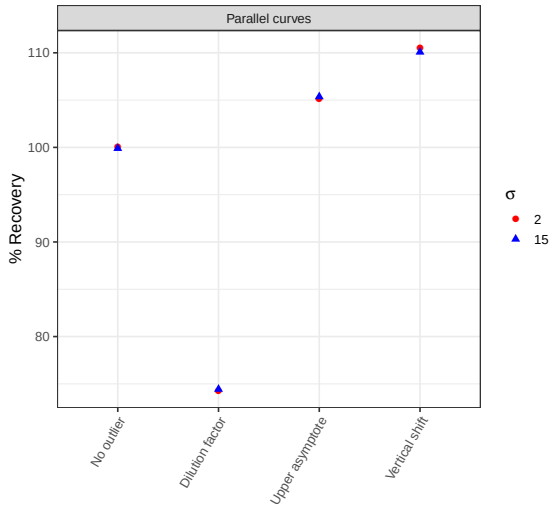


Figure B.12: Whole curve outlier: geometric mean of the recovery when true RP = 200%

Appendix C

Comparison of Outlier Tests for All Outlier Types and Measurement Variabilities

This Appendix presents the supplementary material of Chapter 4. In the chapter, we only present moderate outliers results and, for similarity testing and RP estimation, only results when $\sigma = 2$. This Appendix shows the same results for mild and extreme outliers as well as moderate outliers when $\sigma = 15$.

C.1 True Positive Rates for mild and extreme outliers

Figures C.1 to C.4 present the true positive rates for mild and extreme outliers. In general, the results follow similar pattern to the ones presented in Chapter 4, with ROUT being generally better than other outlier tests. A key difference is that RPI detects more mild outliers than ROUT in most cases.

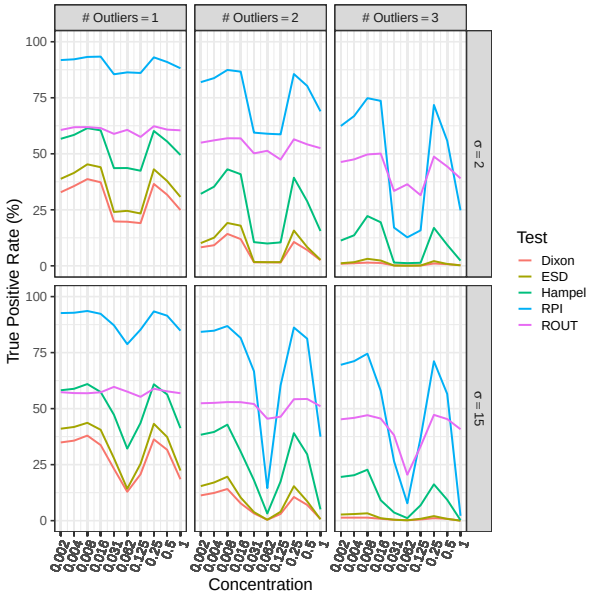


Figure C.1: True Positive Rate at specific concentration points, mild outliers

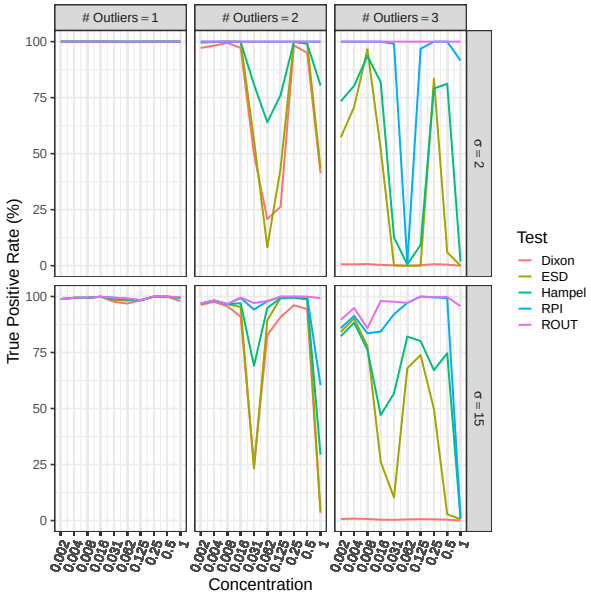


Figure C.2: True Positive Rate at specific concentration points, extreme outliers

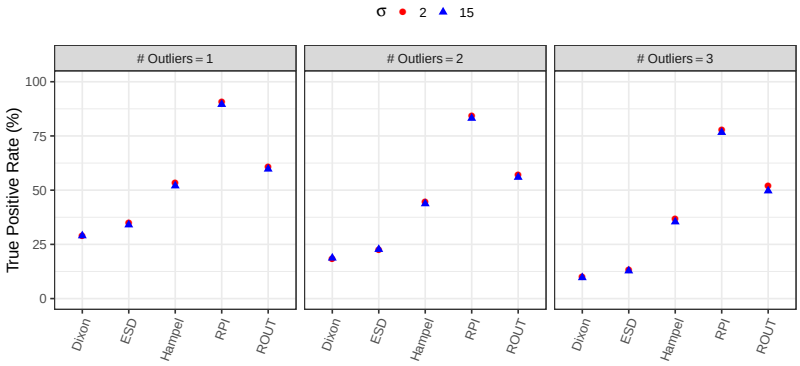


Figure C.3: True Positive Rate across random concentration points for mild outliers

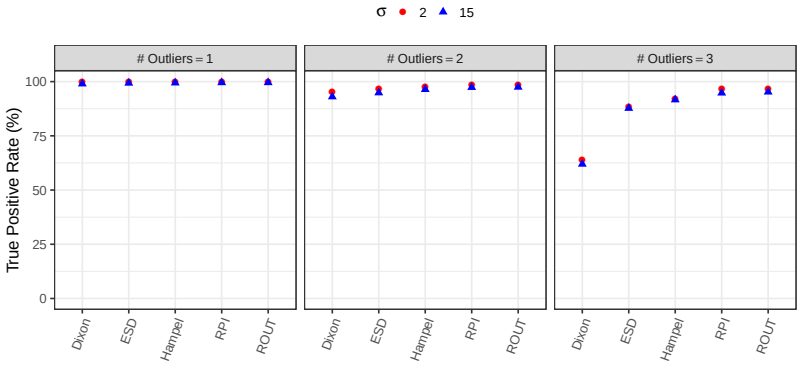


Figure C.4: True Positive Rate across random concentration points for extreme outliers

C.2 Parallelism test results for mild and extreme outliers

Figures C.5 and C.6 present the parallelism test results for mild and extreme outliers. These results follow a similar pattern as the results presented in Chapter 4, but RPI has a better parallelism detection than ROUT in some mild outlier cases.

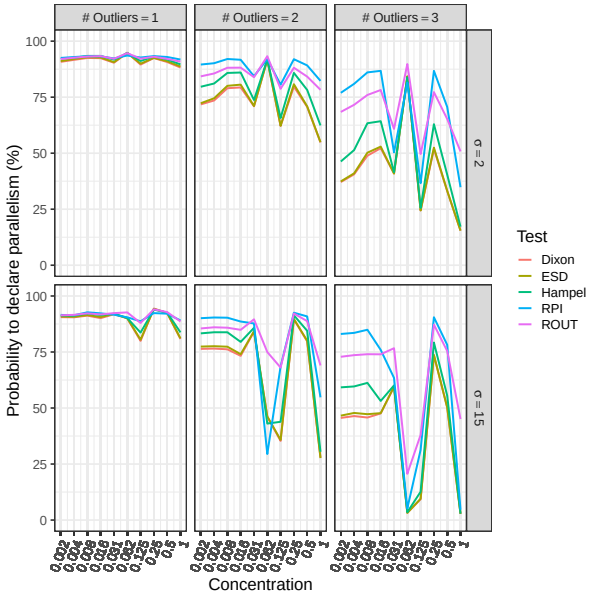


Figure C.5: Probability to declare similarity test after detection and removal of mild outliers

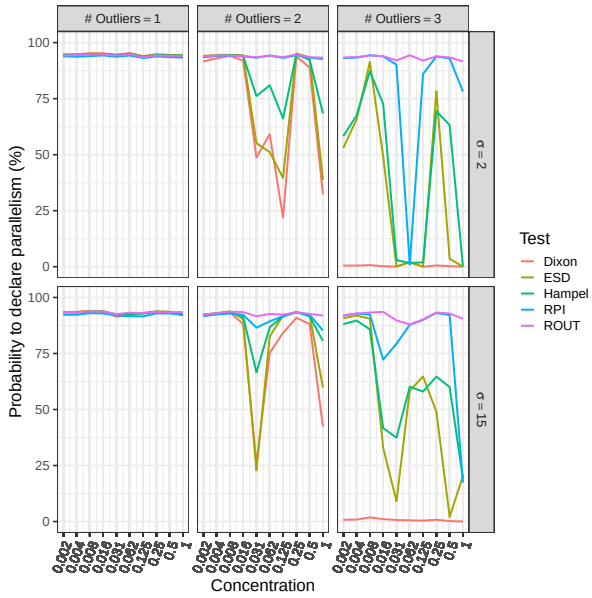


Figure C.6: Probability to declare similarity test after detection and removal of extreme outliers

C.3 Recovery for all outlier types and measurement variabilities

Figures C.7 to C.24 present the % recovery for all outlier types and measurement variabilities. Again, the result pattern is highly similar to the one presented in Chapter 4, with different magnitude by measurement variability and higher recoveries for mild outliers because they are less often detected than moderate and extreme outliers. For tests that fail to detect all extreme outliers, the opposite situation arises, and extreme outliers logically induce a higher error.

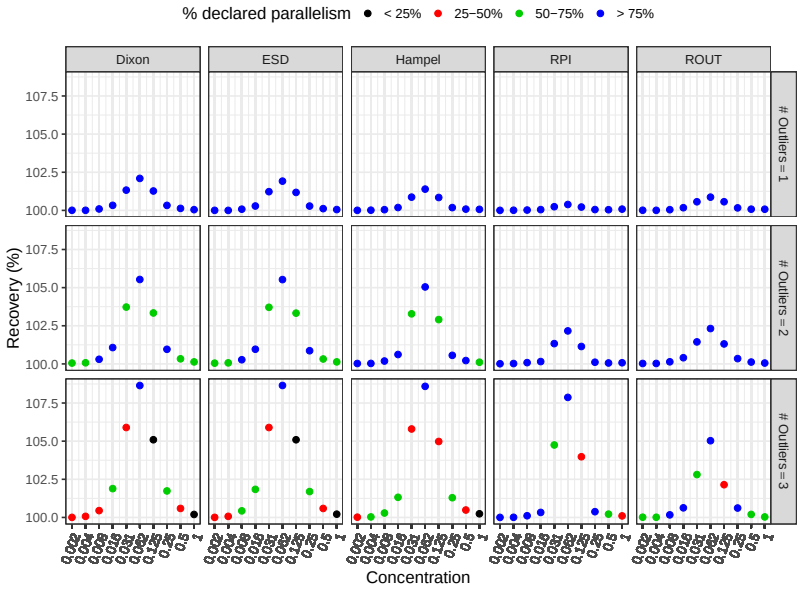


Figure C.7: Geometric Mean %recovery in estimated relative potency (RP) after detection and removal of mild outliers when $\sigma = 2$

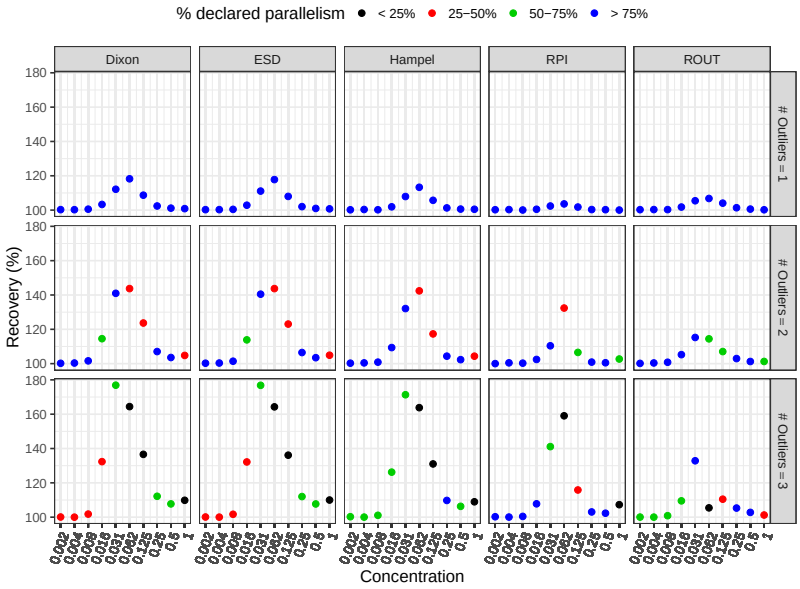


Figure C.8: Geometric Mean %recovery in estimated relative potency (RP) after detection and removal of mild outliers when $\sigma = 15$

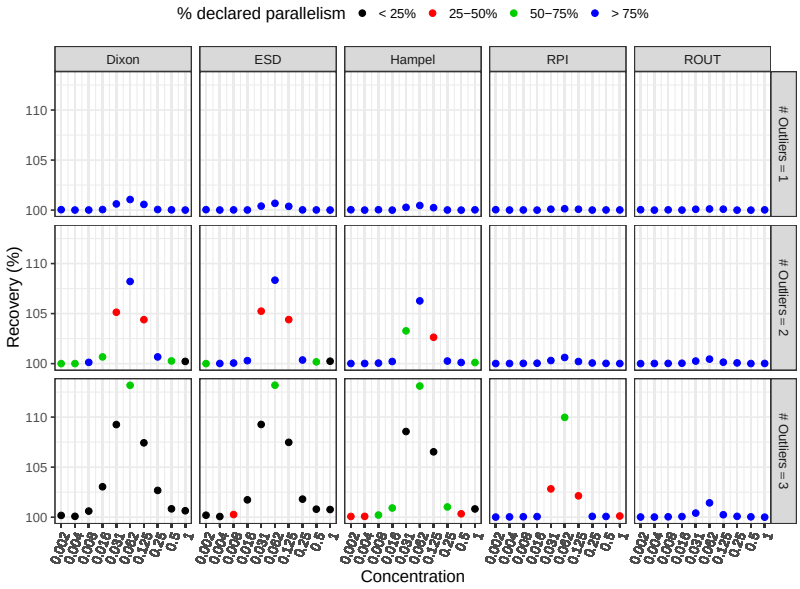


Figure C.9: Geometric Mean %recovery in estimated relative potency (RP) after detection and removal of moderate outliers when $\sigma = 2$

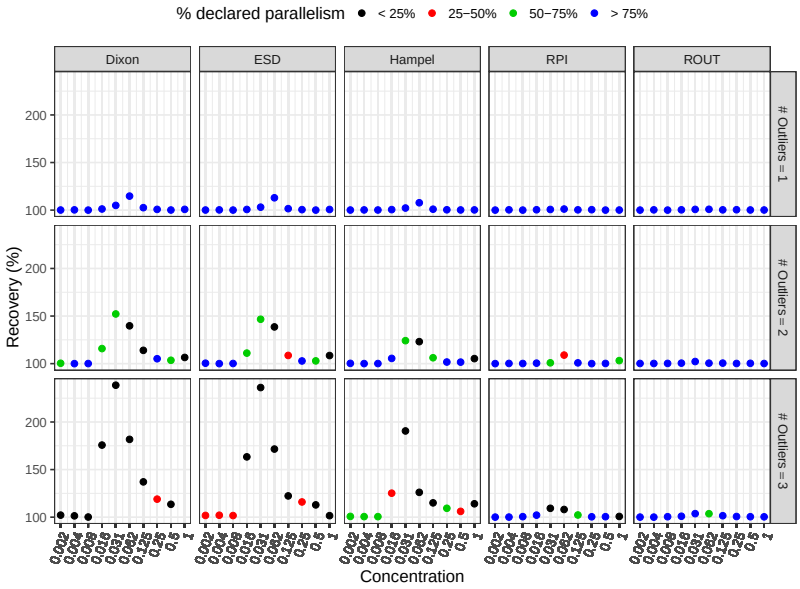


Figure C.10: Geometric Mean %recovery in estimated relative potency (RP) after detection and removal of moderate outliers when $\sigma = 15$

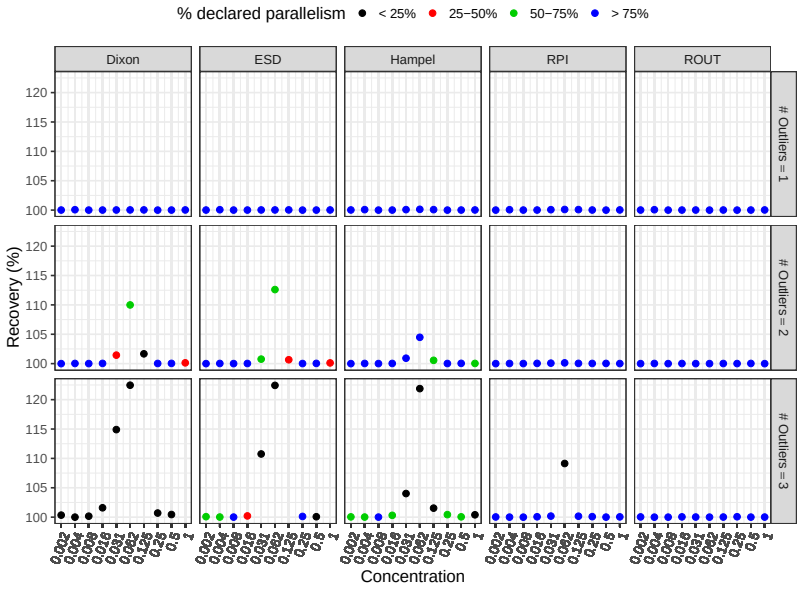


Figure C.11: Geometric Mean %recovery in estimated relative potency (RP) after detection and removal of extreme outliers when $\sigma = 2$

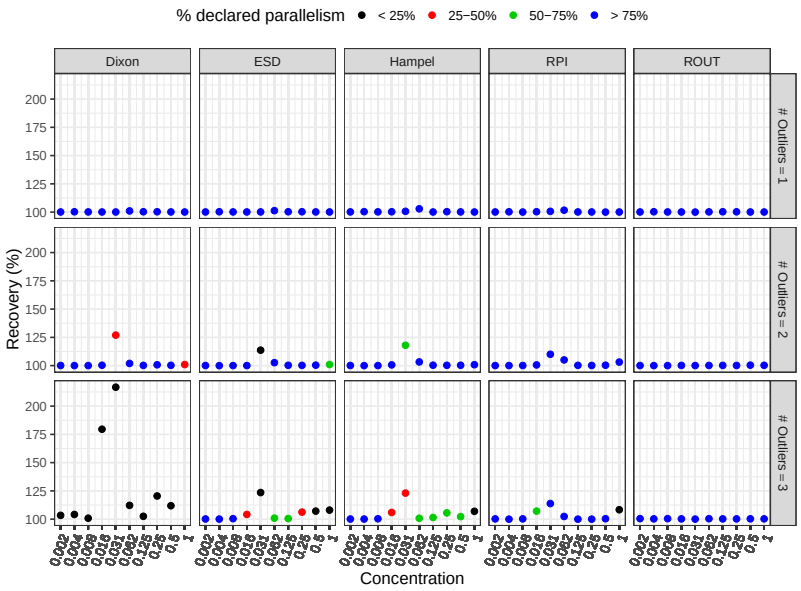


Figure C.12: Geometric Mean %recovery in estimated relative potency (RP) after detection and removal of extreme outliers when $\sigma = 15$

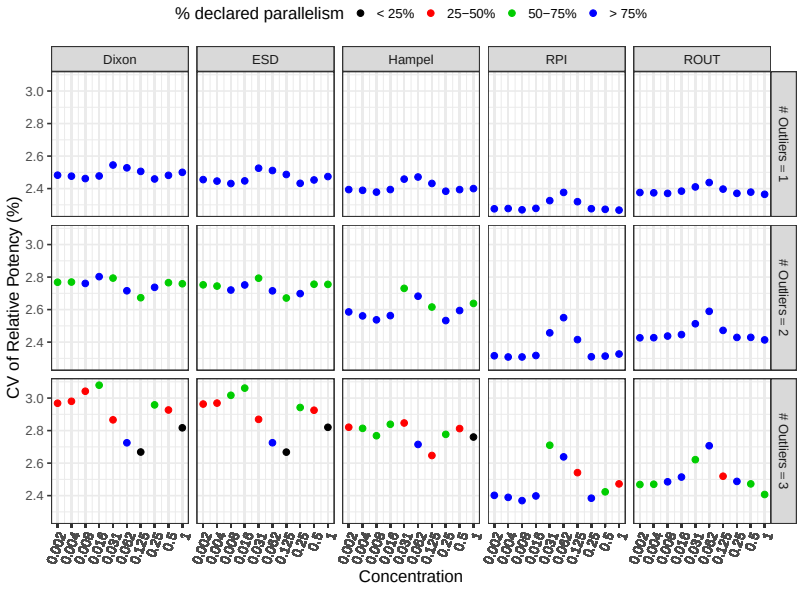


Figure C.13: Median %CV of the estimated relative potency (RP) after detection and removal of mild outliers when $\sigma = 2$

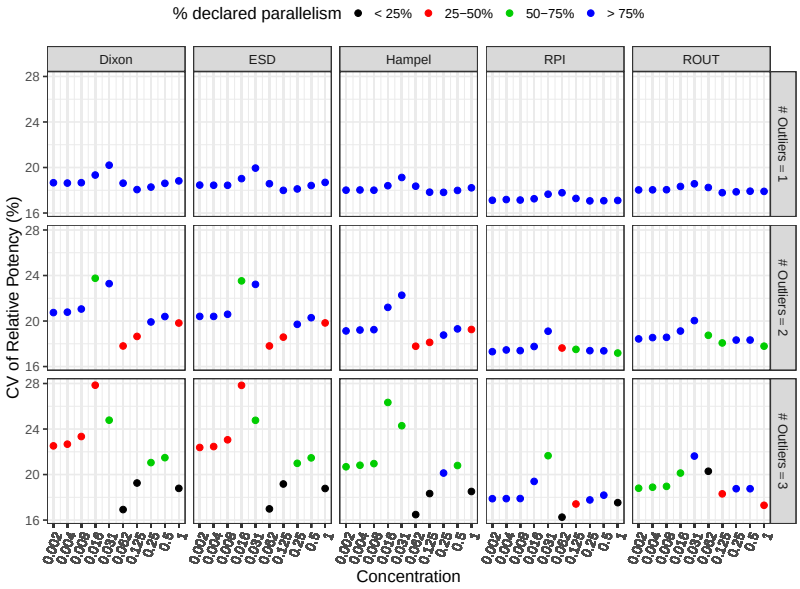


Figure C.14: Median %CV of the estimated relative potency (RP) after detection and removal of mild outliers when $\sigma = 15$

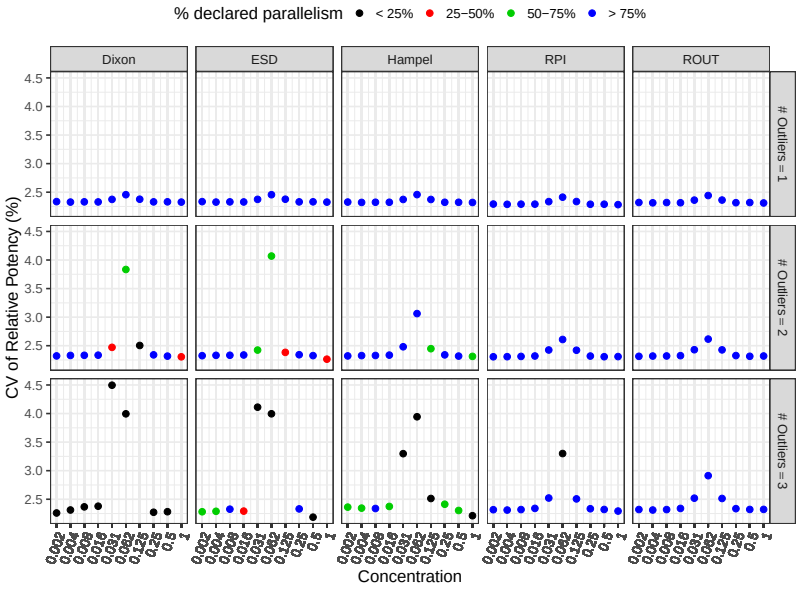


Figure C.15: Median %CV of the estimated relative potency (RP) after detection and removal of extreme outliers when $\sigma = 2$

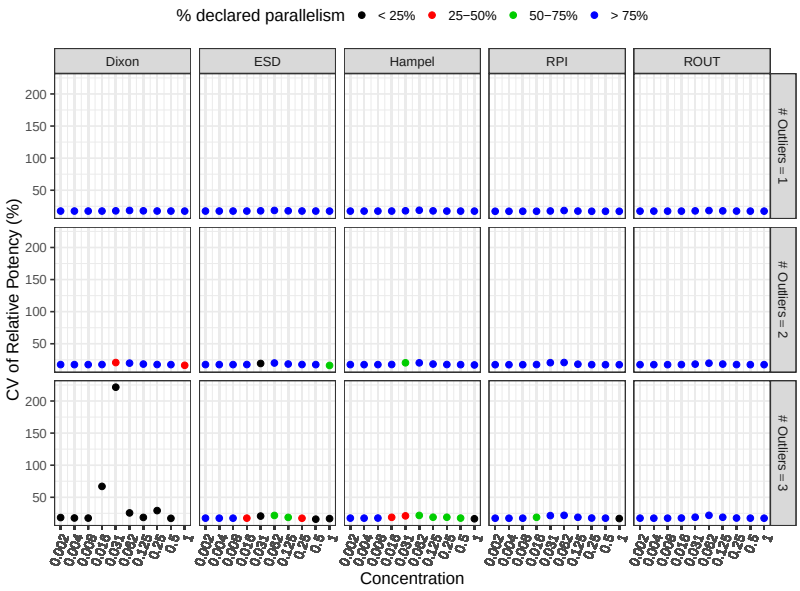


Figure C.16: Median %CV of the estimated relative potency (RP) after detection and removal of extreme outliers when $\sigma = 15$

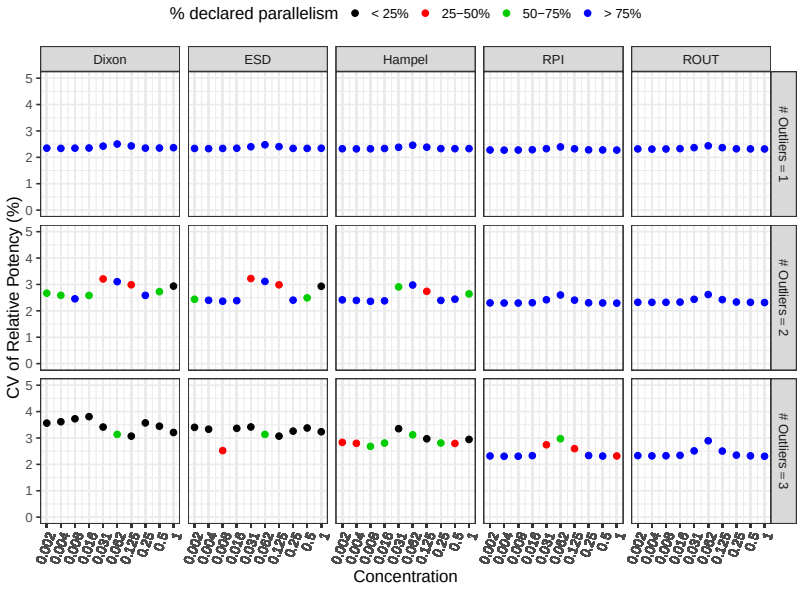


Figure C.17: Median %CV of the estimated relative potency (RP) after detection and removal of moderate outliers when $\sigma = 2$

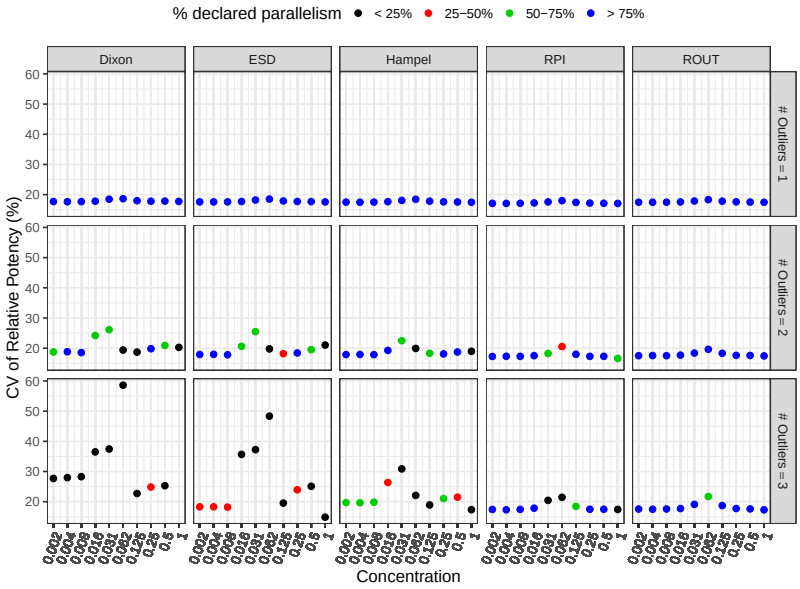


Figure C.18: Median %CV of the estimated relative potency (RP) after detection and removal of moderate outliers when $\sigma = 15$

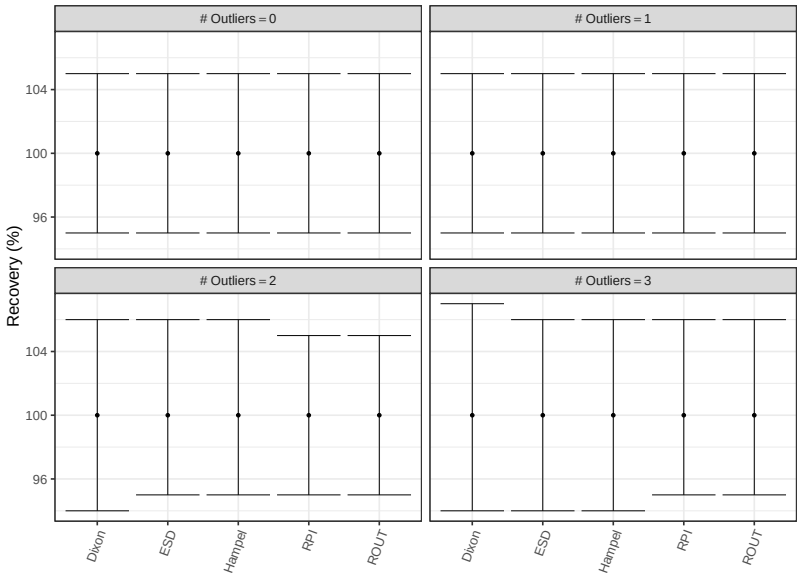


Figure C.19: 95% Monte Carlo coverage interval of estimated recovery after detection and removal of mild outliers at random positions $\sigma = 2$

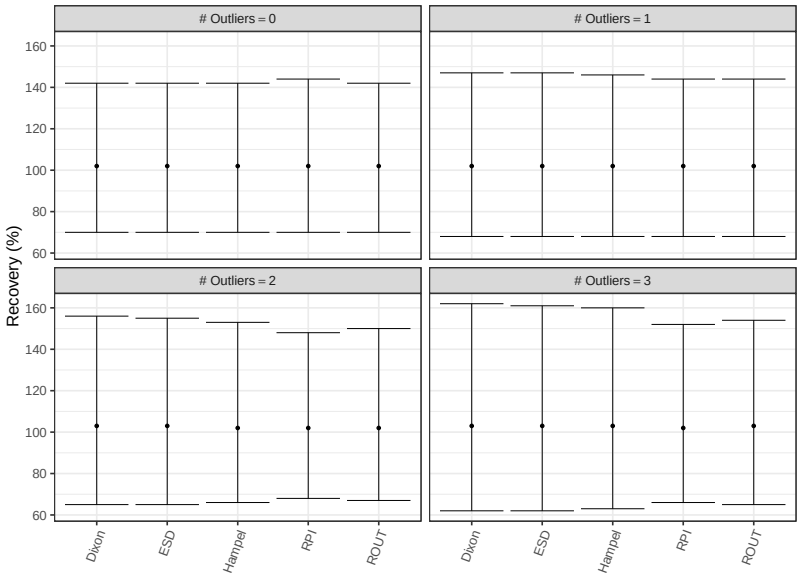


Figure C.20: 95% Monte Carlo coverage interval of estimated recovery after detection and removal of mild outliers at random positions $\sigma = 15$

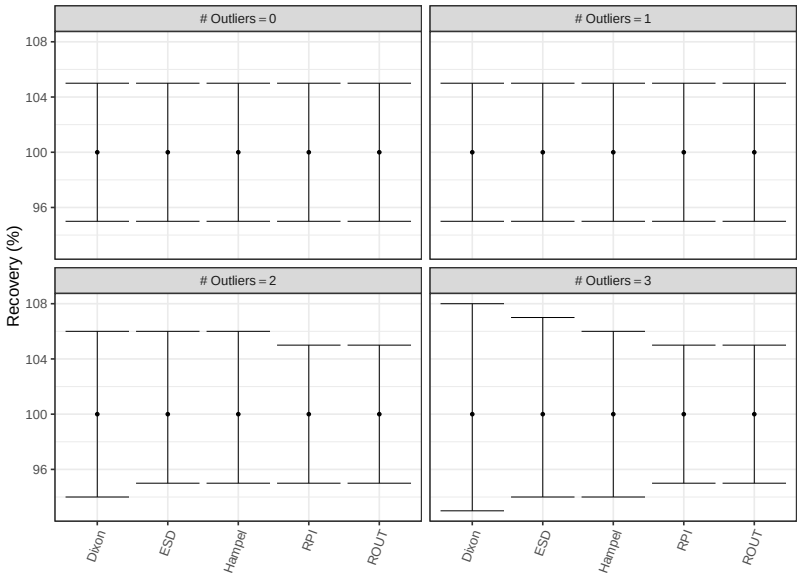


Figure C.21: 95% Monte Carlo coverage interval of estimated recovery after detection and removal of moderate outliers at random positions $\sigma = 2$

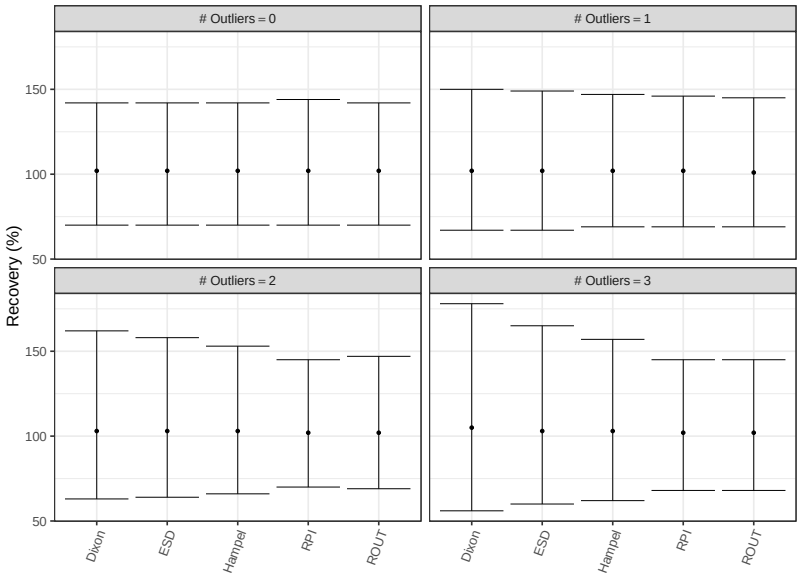


Figure C.22: 95% Monte Carlo coverage interval of estimated recovery after detection and removal of moderate outliers at random positions $\sigma = 15$

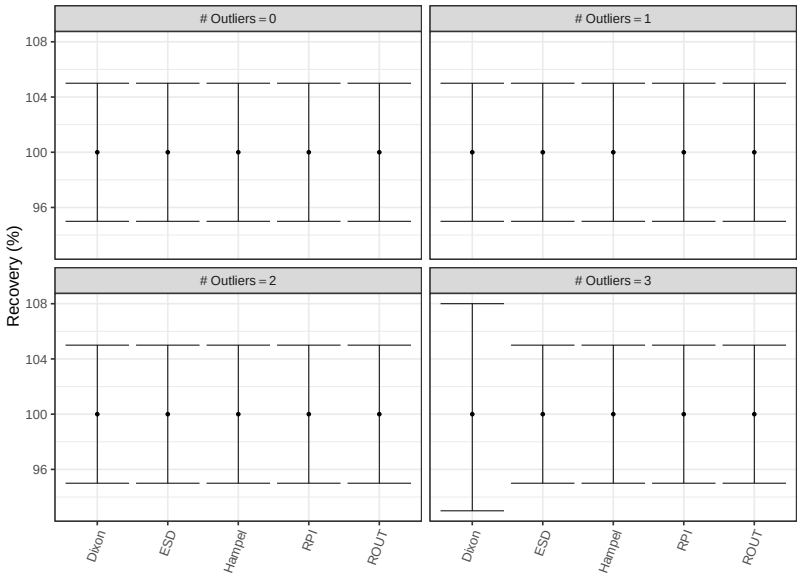


Figure C.23: 95% Monte Carlo coverage interval of estimated recovery after detection and removal of extreme outliers at random positions $\sigma = 2$

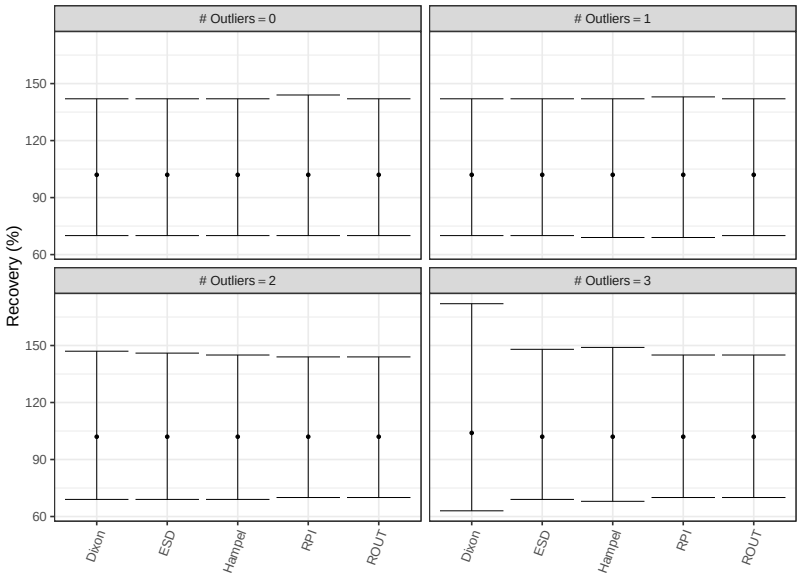


Figure C.24: 95% Monte Carlo coverage interval of estimated recovery after detection and removal of extreme outliers at random positions $\sigma = 15$

This page is intentionally left blank

Appendix D

Error in $\log(RP)$ and Non-Similarity Parameters by Experimental Sesign for all Measurement Variabilities

This Appendix presents the supplementary material to Chapter 5. In the chapter, we only present results for selected simulation parameters when $\sigma = 2\%$ of $yMax_R$.

D.1 Scatterplot Matrices of $\log(RP)$ and non-similarity parameters as a function of selected simulation parameters when $\sigma = 10\%$ of $yMax_R$

Figures D.1 to D.4 present the scatterplot matrices of the log(RP) and non-similarity parameters as a function of selected simulation parameters when $\sigma = 0.10 \times yMax_R$. It appears that they present the same pattern as Figures 5.1 to 5.4, with different magnitudes due to the increased measurement variability.

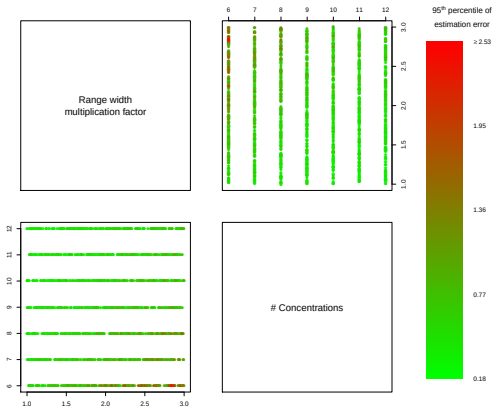


Figure D.1: Scatterplot matrix of the 95^{th} percentile of the absolute difference between the observed and true log(RP) as a function of the log(RP), range width multiplication factor and number of concentrations when $\sigma = 0.10 \times yMax_R$. Light green points represent simulation designs with small estimation error and red points represent simulation designs with high estimation error.

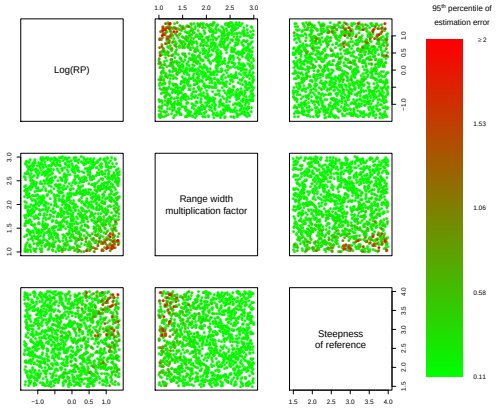


Figure D.2: Scatterplot matrix of the 95^{th} percentile of the absolute difference between the observed and true log(ratio) of upper asymptotes as a function of the log(RP), range width multiplication factor and steepness of reference curve when $\sigma = 0.10 \times yMax_R$. Light green points represent simulation designs with small estimation error and red points represent simulation designs with high estimation error.

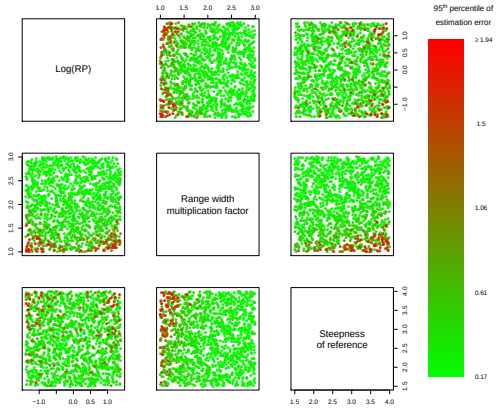


Figure D.3: Scatterplot matrix of the 95^{th} percentile of the absolute difference between the observed and true log(ratio) of asymptote ranges as a function of the log(RP), range width multiplication factor and steepness of reference curve when $\sigma = 0.10 \times yMax_R$. Light green points represent simulation designs with small estimation error and red points represent simulation designs with high estimation error.

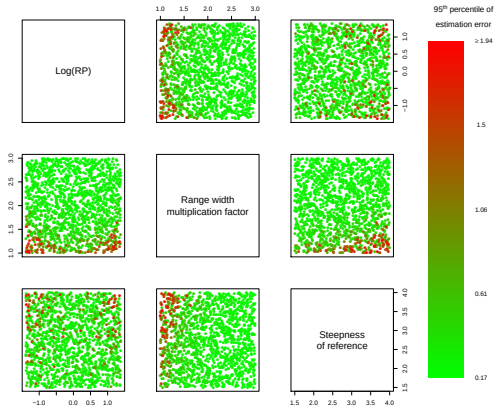


Figure D.4: Scatterplot matrix of the 95^{th} percentile of the absolute difference between the observed and true log(ratio) of slopes as a function of the range width multiplication factor and number of concentrations when $\sigma = 0.10 \times yMax_R$. Light green points represent simulation designs with small estimation error and red points represent simulation designs with high estimation error.

D.2 $\log(RP)$ and non-similarity parameters as a function of all simulation parameters

Figures D.5 and D.6 present 95^{th} percentile of the absolute difference between the observed and true $\log(RP)$ as a function of all simulation parameters, respectively when $\sigma = 0.02 \times yMax_R$ and $\sigma = 0.10 \times yMax_R$. It appears that the quality of estimation of the $\log(RP)$ is mostly affected by the true $\log(RP)$, the range width multiplication factor and the number of concentrations.

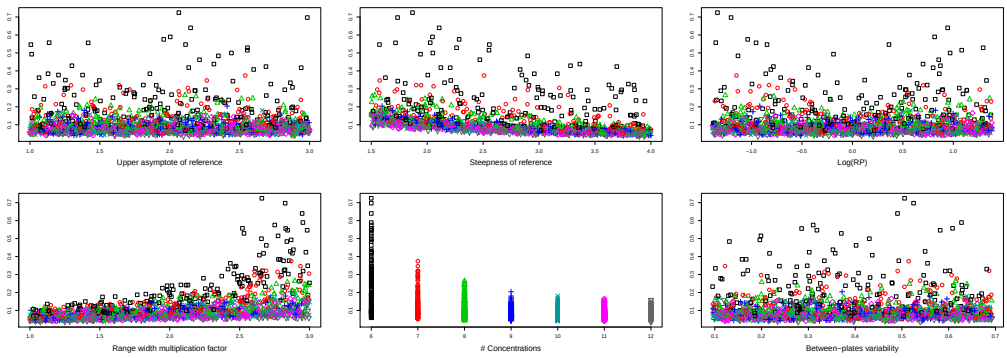


Figure D.5: 95^{th} percentile of the absolute difference between the observed and true log(RP) as a function of all simulation parameters, separately, when $\sigma = 0.02 \times yMax_R$. Different colors represent different number of concentrations per curves.

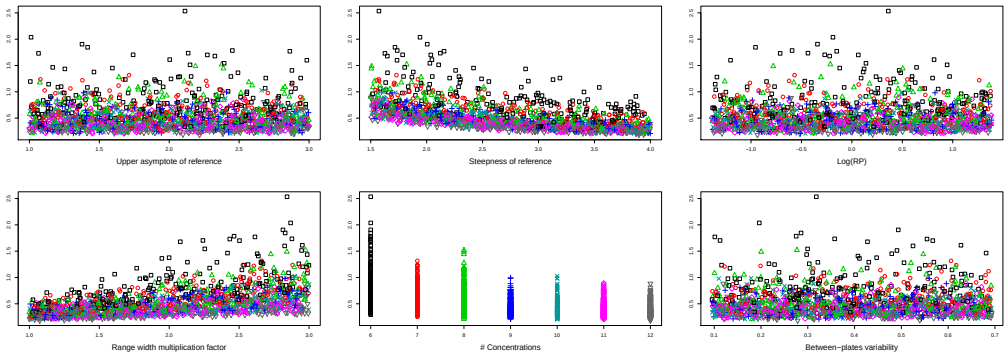


Figure D.6: 95^{th} *percentile of the absolute difference between the observed and true log(RP) as a function of all simulation parameters, separately, when* $\sigma = 0.10 \times yMax_R$. *Different colors represent different number of concentrations per curves.*

Figures D.7 and D.8 present 95^{th} percentile of the absolute difference between the observed and true log(ratio) of upper asymptotes as a function of all simulation parameters, respectively when $\sigma = 0.02 \times yMax_R$ and $\sigma = 0.10 \times yMax_R$. It appears that the quality of estimation of the log(ratio) of upper asymptotes is mostly affected by the true $\log(RP)$, the range width multiplication factor and steepness of the reference curve.

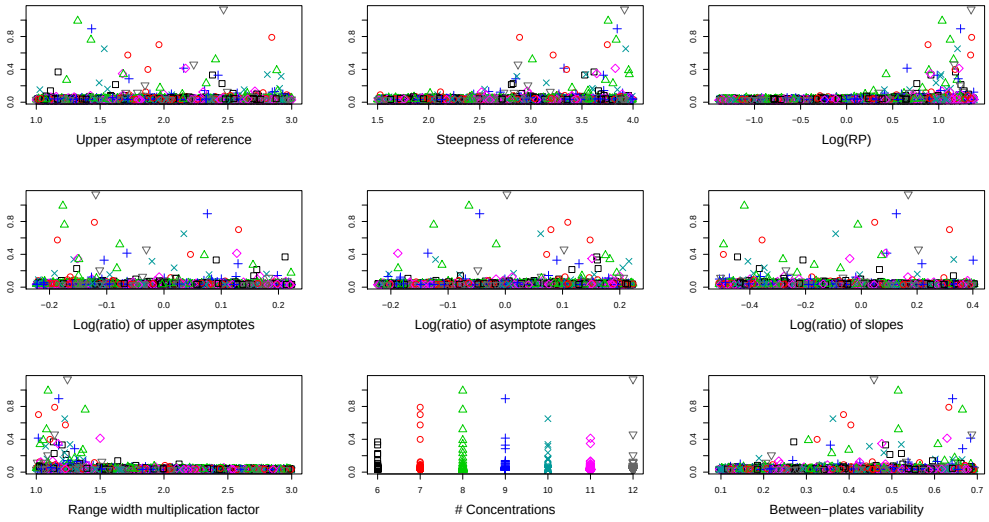


Figure D.7: 95^{th} *percentile of the absolute difference between the observed and true log(ratio) of upper asymptotes as a function of all simulation parameters, separately, when* $\sigma = 0.02 \times yMax_R$*. Different colors represent different number of concentrations per curves.*

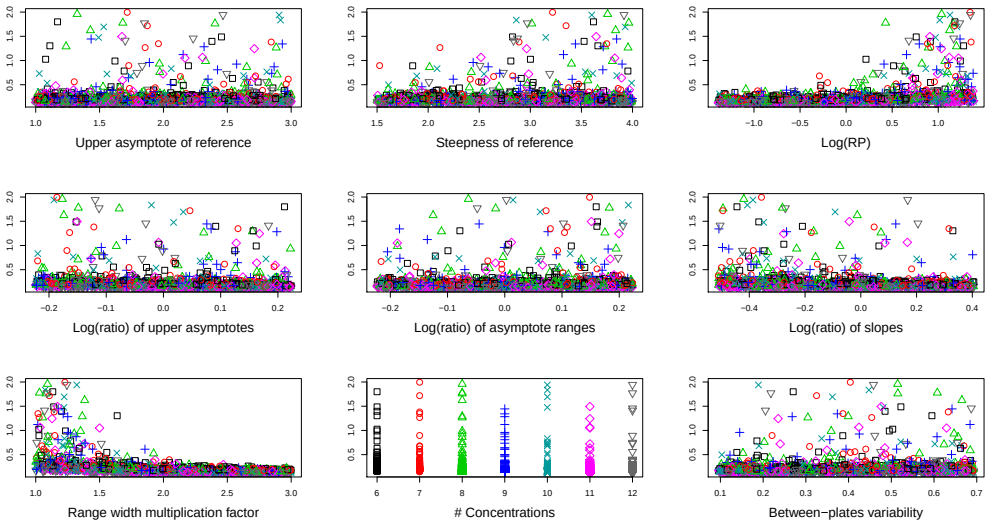


Figure D.8: 95^{th} *percentile of the absolute difference between the observed and true log(ratio) of upper asymptotes as a function of all simulation parameters, separately, when* $\sigma = 0.10 \times yMax_R$. *Different colors represent different number of concentrations per curves.*

Figures D.9 and D.10 present 95^{th} percentile of the absolute difference between the observed and true log(ratio) of asymptote ranges as a function of all simulation parameters, respectively when $\sigma = 0.02 \times yMax_R$ and $\sigma = 0.10 \times yMax_R$. It appears that the quality of estimation of the log(ratio) of asymptote ranges is mostly affected by the true $\log(RP)$, the range width multiplication factor and steepness of the reference curve.

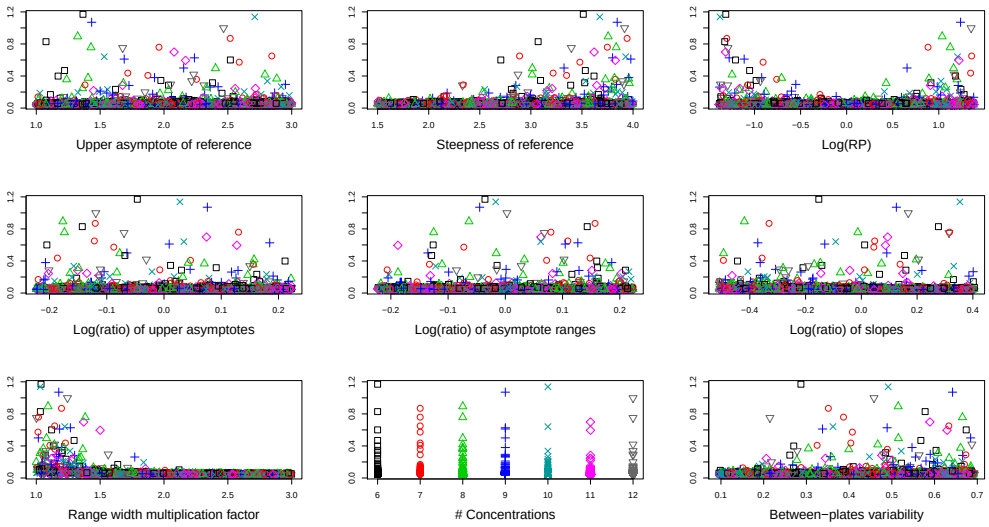


Figure D.9: 95^{th} *percentile of the absolute difference between the observed and true log(ratio) of asymptotes ranges as a function of all simulation parameters, separately, when* $\sigma = 0.02 \times yMax_R$*. Different colors represent different number of concentrations per curves.*

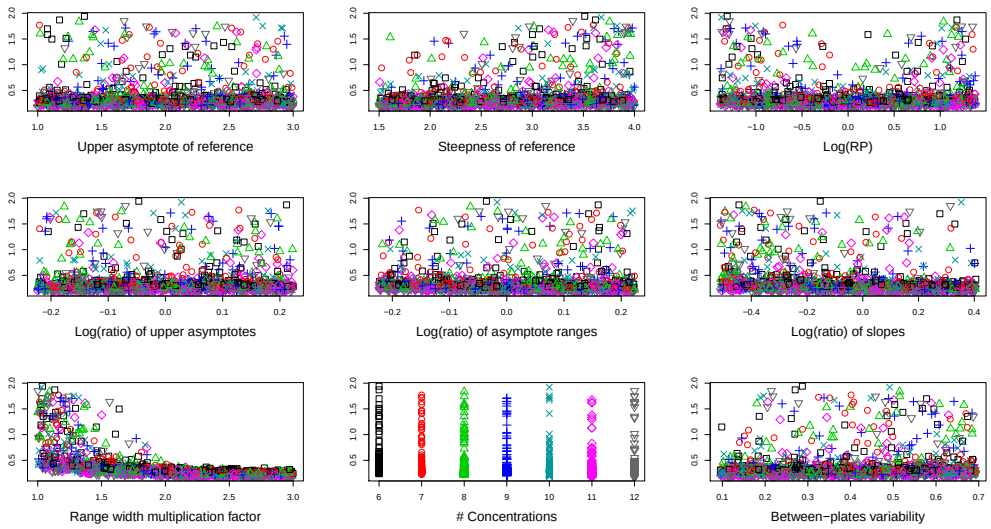


Figure D.10: 95^{th} *percentile of the absolute difference between the observed and true log(ratio) of asymptote ranges as a function of all simulation parameters, separately, when* $\sigma = 0.10 \times yMax_R$*. Different colors represent different number of concentrations per curves.*

Figures D.11 and D.12 present 95^{th} percentile of the absolute difference between the observed and true log(ratio) of slopes as a function of all simulation parameters, respectively when $\sigma = 0.02 \times yMax_R$ and $\sigma = 0.10 \times yMax_R$. It appears that the quality of estimation of the log(ratio) of slopes is mostly affected by the range width multiplication factor and the number of concentrations.

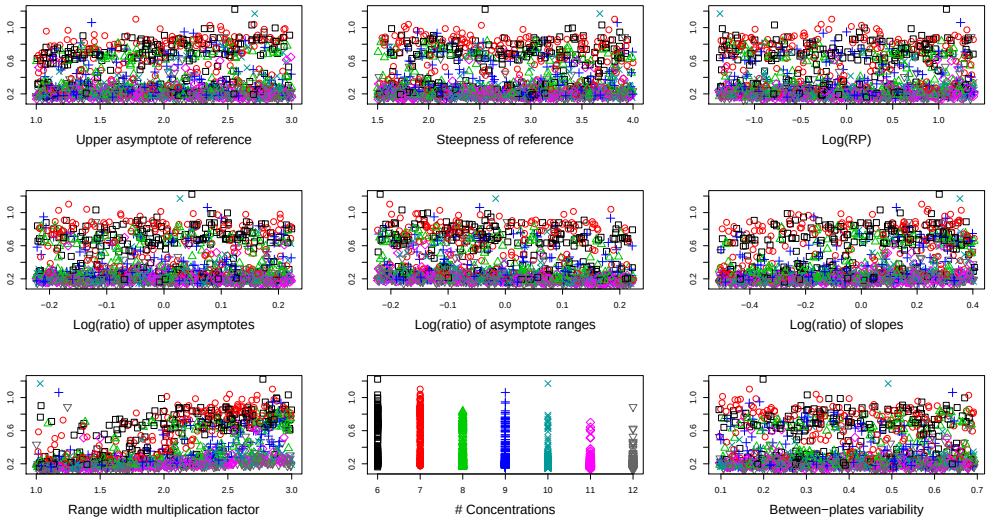


Figure D.11: 95^{th} *percentile of the absolute difference between the observed and true log(ratio) of slopes as a function of all simulation parameters, separately, when* $\sigma = 0.02 \times yMax_R$. *Different colors represent different number of concentrations per curves.*

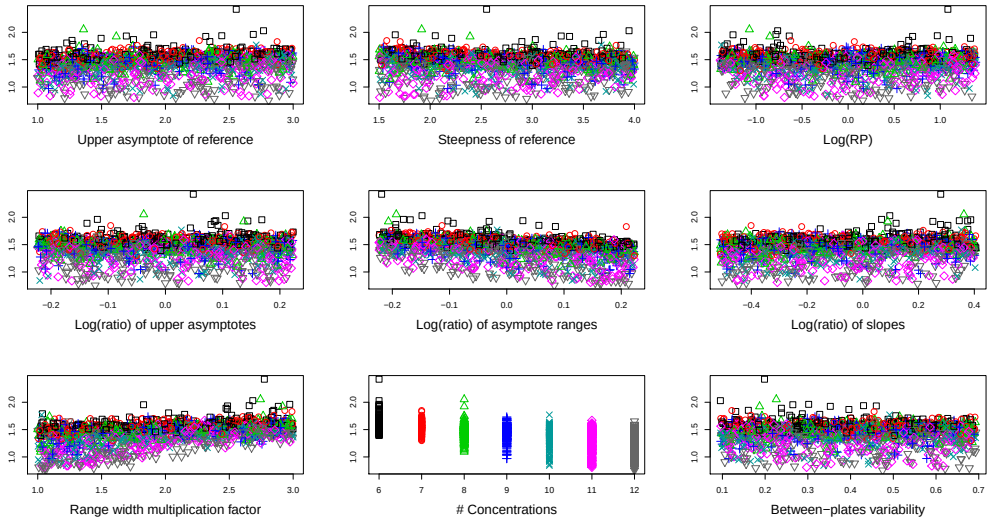


Figure D.12: 95^{th} *percentile of the absolute difference between the observed and true log(ratio) of slopes as a function of all simulation parameters, separately, when* $\sigma = 0.10 \times yMax_R$. *Different colors represent different number of concentrations per curves.*

Figures D.13 and D.14 present Monte Carlo standard deviation of the $RSSE_{nonpar}$, scaled over $yMax_R^2$ as a function of all simulation parameters, respectively when $\sigma = 0.02 \times yMax_R$ and $\sigma = 0.10 \times yMax_R$. It appears that the variability of estimation of the scaled $RSSE_{nonpar}$ is mostly affected by the true log(ratio) of curve parameters, rather than by design aspects.

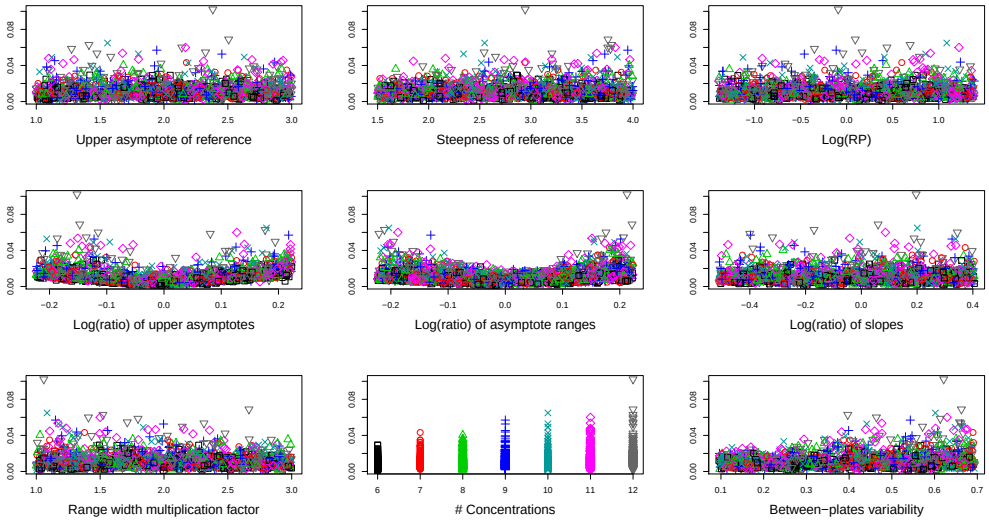


Figure D.13: Monte Carlo standard deviation of the scaled $RSSE_{nonpar}$ as a function of all simulation parameters, separately, when $\sigma = 0.02 \times yMax_R$. Different colors represent different number of concentrations per curves.

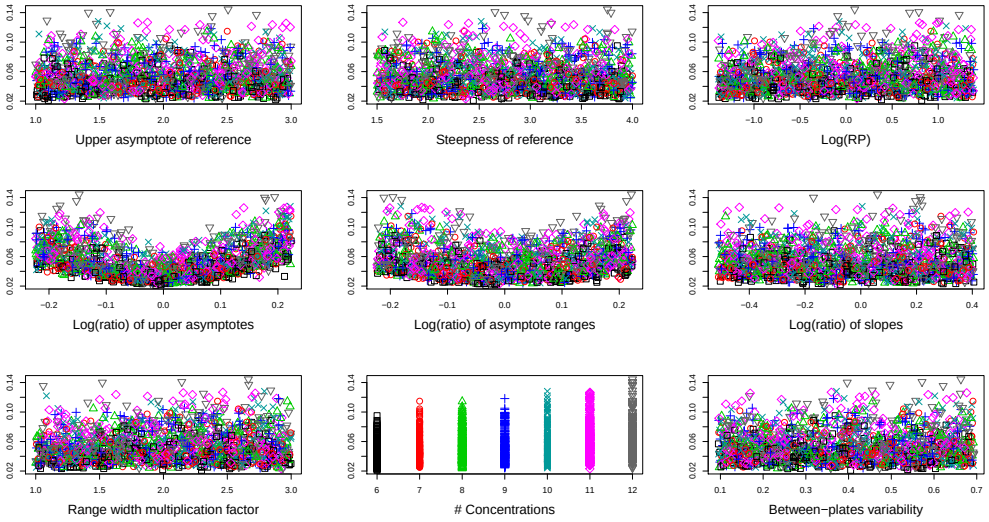


Figure D.14: Monte Carlo standard deviation of the scaled $RSSE_{nonpar}$ as a function of all simulation parameters, separately, when $\sigma = 0.10 \times yMax_R$. Different colors represent different number of concentrations per curves.

Figures D.15 and D.16 present Monte Carlo standard deviation of the maximum departure test statistic as a function of all simulation parameters, respectively when $\sigma = 0.02 \times yMax_R$ and $\sigma = 0.10 \times yMax_R$. It appears that the variability of estimation of the scaled maximum departure is slightly affected by the true $\log(RP)$ and the range width multiplication factor.

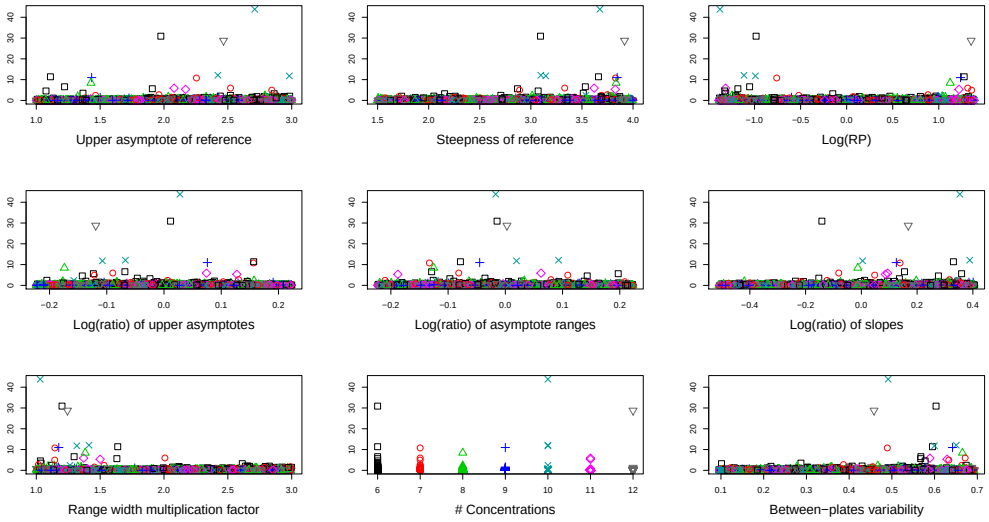


Figure D.15: Monte Carlo standard deviation of the maximum departure test statistic as a function of all simulation parameters, separately, when $\sigma = 0.02 \times yMax_R$*. Different colors represent different number of concentrations per curves.*

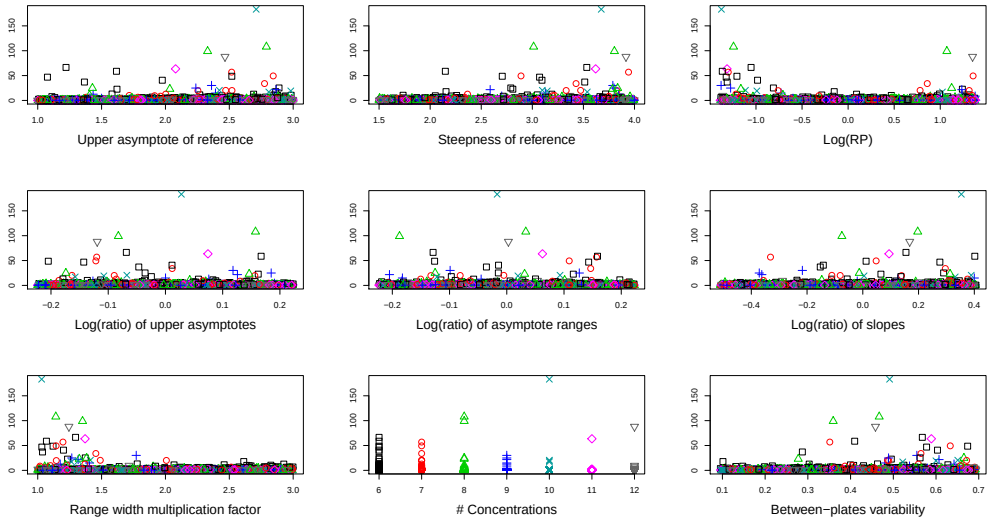


Figure D.16: Monte Carlo standard deviation of the maximum departure test statistic as a function of all simulation parameters, separately, when $\sigma = 0.10 \times yMax_R$*. Different colors represent different number of concentrations per curves.*