Synergy detection: A practical guide to statistical assessment of potential drug combinations

Peer-reviewed author version

Makariadou, Elli; Wang , Xuechen; Hein, Nicholas; DERESA, Negera Wakgari; Mutambanengwe, Kathy; Verbist, Bie & THAS, Olivier (2024) Synergy detection: A practical guide to statistical assessment of potential drug combinations. In: PHARMACEUTICAL STATISTICS,.

DOI: 10.1002/pst.2383
Handle: http://hdl.handle.net/1942/42872

Synergy detection: a practical guide to statistical assessment of potential drug combinations.

Elli Makariadou[1], Xuechen Wang[2*], Nicholas Hein[2], Negera W. Deresa[3], Kathy Mutambanengwe[4], Bie Verbist[1], and Olivier Thas[3,5,6]

[1]Janssen Pharmaceutical Companies of Johnson and Johnson, Beerse, Belgium
[2]Janssen Pharmaceutical Companies of Johnson and Johnson, Spring House, USA
[3]Data Science Institute, I-Biostat, Hasselt University, Hasselt, Belgium
[4]Open Analytics, Antwerpen, Belgium
[5]National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, New South Wales, Australia
[6]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium
[*]Address correspondence to: xwang449@its.jnj.com

Abstract

Combination treatments have been of increasing importance in drug development across therapeutic areas to improve treatment response, minimize the development of resistance, and/or minimize adverse events. Pre-clinical in-vitro combination experiments aim to explore the potential of such drug combinations during drug discovery by comparing the observed effect of the combination with the expected treatment effect under the assumption of no interaction (i.e, null model). This tutorial will address important design aspects of such experiments to allow proper statistical evaluation. Additionally, it will highlight the Biochemically Intuitive Generalized Loewe methodology (BIGL R package available on CRAN) to statistically detect deviations from the expectation under different null models. A clear advantage of the methodology is the quantification of the effect sizes, together with confidence interval while controlling the directional false coverage rate. Finally, a case study will showcase the workflow in analyzing combination experiments.

1 Introduction

Combination therapy, a treatment modality that combines two or more therapeutic agents, is of growing importance in drug development across multiple therapeutic areas. Co-administration of

compounds may be necessary to account for disease complexity and increase efficacy while potentially reducing drug resistance, and/or minimizing adverse events. Consequently, combinations of compounds are routinely screened in pre-clinical in-vitro experiments to identify the most effective drug combinations.

The establishment of a methodology to quantify the presence of synergistic or antagonistic effects is of critical importance. Such an assessment typically relies on the dose-response curves of individual compounds, called monotherapies. Synergy or antagonism is detected when the observed response of a drug combination is different from the expected treatment response under the assumption of no interaction (i.e., the null model) such that the direction of deviation determines synergy or antagonism. The expected treatment responses are derived solely from the monotherapies. Several null models, including Highest Single Agent (HSA) [1], the Bliss Independence Model [2] and the Loewe additivity model [3] have been proposed in the literature without an agreement on the most suitable choice [4]. The models differ on the assumptions of expectation under no interaction and thus, can differ on the conclusions about the detection or degree of synergy/antagonism. For the remainder of this tutorial paper, without loss of generality, we will focus on synergistic effects.

Many software packages, relying on the above concept, often referred to as deviance assessment, have been published. Alternatively, one can perform some parametric modelling of a synergy index [5, 6] which is out of scope for this tutorial paper. Table 1 shows an overview of deviance assessment software packages frequently used in the pharmaceutical industry. To the best of our knowledge the details of the software packages are correct; however, we were unable to test these software packages and we extracted the details from publicly available documentation.

Table 1: Overview of deviance assessment software packages.

				Statistics reported	
Software package	Accessibility	Monotherapy	Null Models	Overall	Individual contributions
SynergyFinder	Free	4PL, LOESS or LM	HSA, Bliss, Loewe, and ZIP	Point estimate (sd)	Point estimate (sd)
SynergyFinder Plus	Free	4PL, LOESS or LM	HSA, Bliss, Loewe, and ZIP	Point estimate p-value	Point estimate CI (normal bootstrap)
MacSynergy™	Free	No	Bliss	Point estimate	Point estimate (sd)
Genedata Screener®	Commercial	4PL, 3PL, 2PL	HSA, Bliss, Loewe	Point estimate	Point estimate
CombeneFit	Free	3PL (b^1=1)	HSA, Bliss, Loewe	Point estimate	Point estimate p-value
Chalice™	Commercial	3PL	HSA, Bliss, Loewe and Boost	Point estimate (sd)	Point estimate (sd)
BIGL	Free	4PL, 3PL, 2PL	HSA, Bliss, Generalized Loewe[2]	Point estimate CI (wild bootstrap) p-value	Point estimate CI (wild bootstrap) p-value

[1] b represents the lower asymptote.

[2] The Generalized Loewe null model allows for partial response of the monotherapy data.

4PL, 3PL, 2PL are abbreviations for four-, three- and two-parameter logistic regression, respectively. LM is an abbreviation for linear model. CI is an abbreviation for confidence interval.

Most software packages start with modelling the monotherapy data using a dose-response relationship (e.g., 4-parameter logistic regression). The modelled monotherapy data are then used to predict the expected treatment effects under a specified null model. MacSynergy™ [7] uses the

observed responses to predict the expected treatment responses. Next, the divergence of the observed responses from the expected treatment responses are summarized in synergy scores as excess responses. For any statistical analysis, it is important to evaluate the point estimate (i.e., excess response) relative to the variability of the data expressed as either a confidence interval (CI) and/or p-value. Genedata Screener®, Chalice™, MacSynergy™, and SynergyFinder [8–10] are reporting the excess responses together with observed standard deviations for the latter three. The excess responses in the individual combination points are further summarized as an overall excess response and is often evaluated using a threshold. Routinely, this threshold is not subjected to any hypothesis testing. Failure to evaluate the excess response(s) relative to either CIs or p-values increases the risk of reporting false synergies (i.e., false positives); the rate of these false positives (error rate) should be controlled. SynergyFinder Plus [11], which is an updated version of the original SynergyFinder, recently added a bootstrapping approach to retrieve CIs around the excess responses to reduce the risk of false positive calls, a procedure similar to the Biochemically Intuitive Generalized Loewe (BIGL) implementation [12]. Combenefit [13], alternatively, performs a one sample T-test on the excess responses, whereas BIGL uses an F-test.

The remainder of this tutorial paper will describe important design aspects of drug combination experiments, describe proper statistical evaluation using the BIGL R package [14], and explain the different assumptions of the underlying null models. The BIGL R package was chosen for its ability to:

1. Incorporate several widely used null models, including HSA, Bliss, and Loewe while allowing for partial response of the monotherapy data.
2. Provide flexibility on the monotherapy dose-response models.
3. Perform statistical testing with error rate control, while relaxing distributional assumptions via bootstrapping.
4. Provide effect size estimates (i.e., excess responses), for each combination point and an average effect size, and their confidence intervals.

Lastly, a case study will be presented to illustrate the use of the BIGL R package and how to interpret the results and visualizations from the BIGL R package output.

2 Experimental design

Prior to performing any combination experiment involving two or more monotherapies, experiments should be performed in which the monotherapies are profiled under the same conditions of the combination experiment (e.g., assay, cell line, E:T ratio, incubation time, etc). Profiling the monotherapy is in the form of a dose-response relationship which describes the magnitude of response as a function of dose. This dose-response relationship can be described by dose-response curves and can be mathematically modeled using either a 4-parameter logistic (4PL), 3PL, or 2PL regression model [15].

The 4PL regression model is defined using four parameters that are related to the graphical properties of the curve, i.e., lower asymptote (b), upper asymptote (m), inflection point (EC_{50}), and hill slope (h). These four parameters can be used to express the magnitude of response f, given dose, d, as

$$f(d) = b + \frac{(m-b)}{1 + \left(\frac{\mathrm{EC}_{50}}{d}\right)^{|h|}} \tag{1}$$

Parameters of the 4PL regression model can be fixed. For example, for a particular assay, the absence of any compound elicits no response. One may then fix the lower asymptote of the 4PL regression model to zero. Often, the 4PL regression model with fixed lower or upper asymptote is referred to a 3PL regression model. Fixing an additional parameter (e.g., hill slope) would create a 2PL regression model.

Accurate estimation of the 4PL model parameters is paramount since prediction of the combined treatment responses, assuming no interaction, are estimated using the 4PL monotherapy model parameters. Accurate estimation of these parameters is dependent on the doses selected and the magnitude of responses they elicit. The entire sigmoidal pattern of the monotherapy curve should be covered with the selected doses. A recommendation of the ideal spread of doses is provided in Figure 1, with 2 points at each of the asymptotes and 3 points on the linear part of the curve. Depending on the hill slope of the monotherapy curve, a certain dilution series should be chosen to follow the above recommendation, which is in line with published guidelines [16].

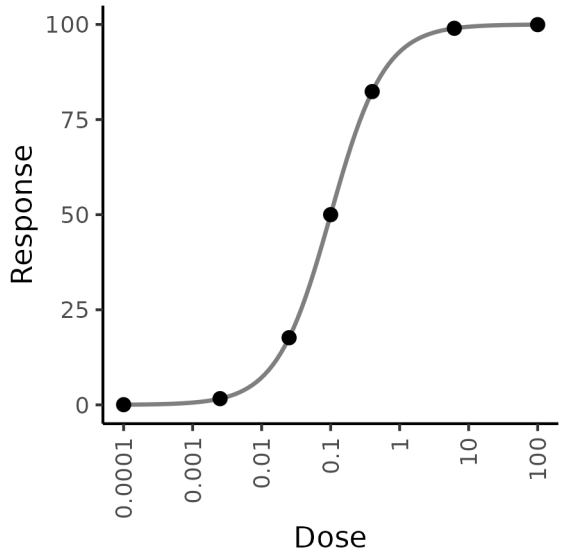


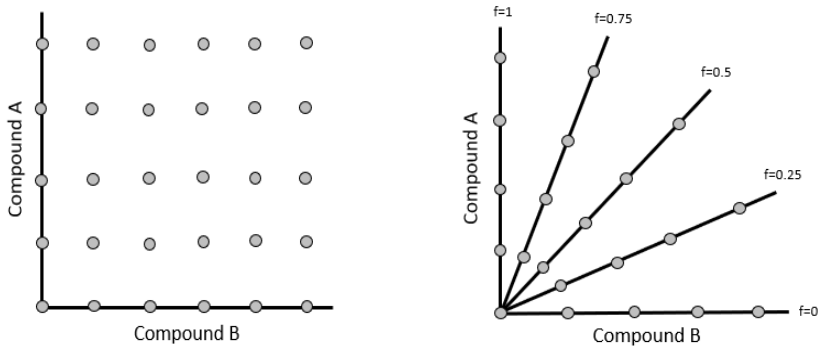
Figure 1: Ideal dilution series with 7 doses.

After choosing the desired doses of the monotherapies, the combination experiment needs to be set up, with two common designs being applied, the ray (dose gradients) and the checkerboard (factorial) designs. The ray design uses fixed ratios of doses while the checkerboard crosses two sets of doses. Examples are given in Figure 2a and 2b for the checkerboard and ray design, respectively.

For the ray design, it is typical to fix the dose titration of one of the monotherapies and multiply the concentrations of the fixed doses by a scalar to get desired ratio for a ray. As such, ray designs tend to have the same number of titrations for each monotherapy and require the scientist to create additional dilutions of one of the monotherapies for each specified ray. Ray designs can be an efficient use of resources and exploration of a set of ratios; however, prior knowledge is often required for proper selection of the rays (drug ratios).

The checkerboard experimental design is a more comprehensive design when prior knowledge is limited. The checkerboard design is a factorial design in which the doses of the monotherapies are crossed with each other. As such, the checkerboard design explores many ratios; however, the number of titrations for each ratio is limited (due to design) excluding the 1:1 ratio (i.e., typically the diagonal of the checkerboard). Furthermore, only a single set of dose titrations for each monotherapy is required; however, the checkerboard design tends to require more plate real estate than the ray design. Lastly, the checkerboard is limited to pairs of monotherapies whereas the ray design can more easily be scaled to triple-, quadruple-, etc., drug combinations.

In summary, we recommend the checkerboard design for the discovery phase of drug development when prior knowledge of biological mechanism is limited, as it's easier and more convenient requiring just one set of dose titrations for each monotherapy. However, when prior knowledge is available, the ray design is highly valuable, offering targeted insights. Additionally, the ray design allows dose response curves to be fit and compared to each ray at the expense of additional laboratory labor, i.e., each ray requires a unique set of dose titrations.



(a) (b)

Figure 2: Drug combination designs. The dots indicate the tested doses of compounds A and B (a) Checkerboard design. (b) Ray design.

Please note that as with any in-vitro experiment, sufficient control wells should be included to access the quality of the assay. This topic is beyond the scope of this tutorial, and we recommend using established guidelines within your organization.

Regardless of the chosen design, the BIGL methodology provides a harmonized framework to analyze drug combination experiments through statistical evaluation of the differences between the expected and observed responses. To estimate variability, replicates of both the monotherapy and

the combination data are required. We currently recommend a minimum of 4 replicates to have an accurate estimation of the observed variability and sufficient power to detect excess responses (supportive information). Furthermore, if the experimental condition requires evaluation of donor effects (e.g., biologics), we recommend against using donors as replicates due to the increased variability. Instead, we recommend replicates within donor and analyzing the donors separately.

3 Methods

Since the BIGL methodology can be categorized as a deviance assessment methodology, a general workflow to detect synergy can be followed:

1. Fit the dose-response curves for the monotherapies.
2. Predict the responses for the combinations using the chosen null model.
3. Estimate the effect sizes, together with CIs.

Fitting monotherapies

The BIGL R package offers flexibility in how the monotherapy dose-response relationships are modelled. The default implementation models the monotherapies using 4PL regression models with a shared asymptote. Depending on whether the monotherapy dose-response curves are increasing or decreasing, either a common lower asymptote or upper asymptote, respectively, is assumed for both drugs. The BIGL R package allows the lower asymptote, upper asymptote, and/or hill slope parameter to be set to a fixed value reducing the 4PL to either a 3PL or 2PL regression model. Setting a parameter to a fixed value must be done with caution and based on biological understanding. Additionally, the BIGL R packages allows linear constraints on the 4PL parameters to facilitate the monotherapy fitting. For instance, if the dose-response curve is decreasing, the minimum estimation could be constrained to be above 0 for an improved biological interpretation.

Predicting expected combination response

The BIGL R package integrates three popular null models, each assumes its own underlying mechanism, which characterizes the no interaction or the expected outcome under a combination of drugs. The HSA quantifies the degree of synergy as the excess over the maximum monotherapy response. Bliss independence assumes the drugs acted independently and synergy is evaluated as the excess of the multiplicative effects of the single drugs. Loewe instead, is a dose-effect based model, calculating the additive (i.e., no interaction) effects as if the single drugs where exchangeable (see Table 2) resulting in an additive effect when the drug is combined with itself. It's worth noting that the true mechanism is often unknown and the described null models are often a simplification of the underlying biology. Therefore, the choice of a particular null model should be informed by domain expertise or results should be compared across null models. Implicitly, the null models assume the same maximal responses of the two monotherapies used in the combination. This assumption is often violated in practice. The BIGL R package relaxes this assumption through the use of the

Generalized Loewe null model. Details of this approach are out of the scope of this tutorial but we refer to Van der Borght 2017 and Thas 2022 for more information and additional alternatives. [12, 17].

Table 2: Overview of the null models.

Null model	Assumption	Formula
HSA	Expected effect is highest effect of monotherapies	$f_{12}(d_1, d_2) = max(f_1(d_1), f_2(d_2))$
Bliss independence	Drugs' effects do not interfere with one another	$f_{12}(d_1, d_2) = f_1(d_1) + f_2(d_2) - f_1(d_1)f_2(d_2)$
Loewe	Two compounds have same mode of action	$\frac{d_1}{f_1^{-1}(f_{12})} + \frac{d_2}{f_2^{-1}(f_{12})} = 1$

In Table 2, $f_{12}(d_1, d_2)$ or f_{12} represents the response at the combination of dose d_1 for drug 1 and dose d_2 for drug 2. $f_i(d_i)$ is the response at dose d_i for drug i. $f_i^{-1}(x)$ is the inverse function of $f_i(d_i)$. It represents the dose of drug i that will produce a response of x.

Estimating effect sizes

The next step is the estimation of the effect sizes and their standard errors. The monotherapy data will be referred to as the on-axis points and the combination data will be referred to as the off-axis points. The effect size at off-axis point $i = 1, \ldots, N$ is defined as the average deviation from the expected treatment response under a null model,

$$E_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (R_{ij} - \hat{R}_i) = \bar{R}_i - \hat{R}_i,$$

where R_{ij} represents the observed response at off-axis point i for replicate j, n_i represents the number of replicates at point i, $\bar{R}_i$ is the sample mean of the observed responses at point i, $\hat{R}_i$ is the estimated treatment response under the chosen null model at off-axis point i, and N is the total number of off-axis points. An overall effect size of synergy which is also named single effect measure is defined as $\overline{E} = \frac{1}{N} \sum_{i=1}^{N} E_i$. With $\mathbf{E}^t = (E_1, \ldots, E_N)$ the vector of all effect size estimates, the variance-covariance matrix of $\mathbf{E}$ can be written as[12, 17]

$$\mathbf{\Sigma} = \text{Var}(\mathbf{E}) = \sigma_0^2 \mathbf{C}_p + \sigma_1^2 \mathbf{D} \tag{2}$$

where σ_0^2 is the residual variance of the on-axis responses, σ_1^2 is the residual variance of the off-axis responses, $\mathbf{D}$ is a diagonal matrix with elements $1/n_i$ $(i = 1, \ldots, N)$, and $\mathbf{C}_p$ is the correlation matrix of the expected treatment responses $(\hat{R}_1, \ldots, \hat{R}_N)$. Specifically, σ_0^2 is estimated as the mean squared error of the 4PL regression model fitting to the monotherapy data. And σ_1^2 is estimated as the average variance of each off-axis point $\frac{1}{N} \sum_i \sum_j A_{ij}^2$ such that $A_{ij} = R_{ij} - \bar{R}_i$. If the assumption of constant variance at the off-axis points does not hold, the variance σ_1^2 in (2) can be replaced by a diagonal matrix with model-based variances on the diagonal positions[12]. The correlation matrix

$\mathbf{C}_p$ is estimated by means of a bootstrap procedure. We outline the bootstrap procedure here in some detail, because it is also needed in the next section.

The estimation of $\mathbf{C}_p$ is as follows:

1. Construct a bootstrap sample of on-axis observations. Specifically, resample the residuals from the 4PL monotherapy regression with replacement. Add these resampled residuals to the fitted values of the 4PL regression model creating a bootstrap sample. Using the bootstrap sample, re-fit the monotherapeutic dose-response curves.

2. Based on the re-fitted dose-response curves, new estimates of the expected treatment responses under the null model are computed for each off-axis point i: $\hat{R}_i^b$ for bootstrap replicate b.

3. Repeat steps 1 and 2 many times (e.g., 1000 times, $B = 1000$).

4. Calculate the sample correlation matrix of the B vectors $(\hat{R}_1^b, \ldots, \hat{R}_N^b)$, denoted by $\hat{\mathbf{C}}_p$.

The sample correlation matrix $\hat{\mathbf{C}}_p$, is an estimator of $\mathbf{C}_p$. The estimator of the covariance matrix $\mathbf{\Sigma}$ can now be written as $\hat{\mathbf{\Sigma}} = \hat{\sigma}_0^2\hat{\mathbf{C}}_p + \hat{\sigma}_1^2\mathbf{D}$. The square roots of its diagonal elements (denoted by s_i) are estimates of the standard errors of the effect sizes. The variance of the overall effect size is given by $\mathrm{Var}(\overline{E}) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} E_i\right)$ and can be estimated as $\frac{1}{N^2}\mathbf{1}^t\hat{\mathbf{\Sigma}}\mathbf{1}$.

Controlling the directional false coverage rate

The original BIGL methodology [17] framed synergy detection in a classical multiple hypothesis testing paradigm, by constructing hypothesis tests based on the effect size estimates, E_i, and their standard errors, s_i, and by controlling the familywise error rate (FWER) at some nominal level. Despite the correctness of this procedure and a positive empirical evaluation[12] we have replaced the hypothesis testing method with a procedure that makes use of simultaneous confidence intervals to control the *directional false coverage rate* (dFCR). Before defining the dFCR, we give two drawbacks of the original approach: (1) controlling the FWER at 5% results in a very conservative detection method (small sensitivity); (2) the results from this testing procedure do not always agree with what would be concluded if confidence intervals were used instead. The first issue could have been resolved by controlling e.g. the false discovery rate (FDR) instead of the FWER, and a solution to the second problem could have been found in aligning the test and CI procedures. However, instead we have chosen to develop a procedure that controls the dFCR.

First, we formally describe a generic method for synergy/antagonism detection based on confidence intervals. Let $[L_i, U_i]$ denote a confidence interval for the effect size at off-axis point i. If $0 \notin [L_i, U_i]$, then we conclude that there is a synergistic or antagonistic effect. We will use the notation τ_i for the true effect size at off-axis point i (i.e., E_i is an estimate of τ_i).

The original definition of the FCR[18] can then be formulated as

$$\mathrm{FCR} = \mathrm{E}\left(\frac{F}{m}\right), \tag{3}$$

where m is the number of off-axis points, and F is the number of intervals among these m points, that do not cover the true effect size τ_i:

$$F = \#\left\{i : \tau_i \notin [L_i, U_i]\right\}.$$

This FCR makes sense in *selective inference*, i.e. statistical inference after the data-driven selection of a subset of parameters. However, in our context we want all confidence intervals to be interpretable, we therefore change the definition from a conditional to a marginal interpretation. Based on very early ideas of the concept of the FDR [19], we further adapt the definition towards a directional FCR. Equation (3) still applies, but now with

$$F = \#\left\{i : \tau_i \notin [L_i, U_i] \text{ and } d(L_i, U_i) \neq \text{sign}(\tau_i)\right\} \tag{4}$$

where $\text{sign}(\tau)$ equals 1 if $\tau > 0$, 0 if $\tau = 0$ and -1 if $\tau < 0$, and $d(L, U)$ equals 1 if $L, U > 0$, 0 if $L < 0$ and $U > 0$ and -1 if $U, L < 0$. In other words, F counts the number of off-axis points for which the conclusion (synergistic / antagonistic / no-effect) is wrong. This is illustrated in Figure 3. Thus, if the dFCR is controlled at 10%, then, on average, 90% of the CIs either contain the true effect size, or at least these CIs result in a correct (directional) conclusion.

Figure 3: Illustration of dFCR. The vertical lines represent confidence intervals and the dots are the true effect sizes τ_i. Left: intervals that contribute to F in the definition of the dFCR. Right: intervals that do not contribute to F.

For controlling the dFCR at a nominal level α, the lower and upper bounds L_i and U_i can be found by means of the bootstrap procedure that was described earlier, but with two additional steps:

1. At each off-axis point i, a Wild bootstrap sample of n_i responses is obtained as $R^b_{ij} = \bar{R}_i + \nu^b_{ij} A_{ij}$, where ν^b_{ij} is randomly sampled from a distribution with stochastic representation

$$\left(\delta_1 + \frac{V_1}{\sqrt{2}}\right)\left(\delta_2 + \frac{V_2}{\sqrt{2}}\right) - \delta_1\delta_2$$

with V_1 and V_2 two independent standard normal random variables, $\delta_1 = 0.5(\sqrt{17/6} + \sqrt{1/6})$ and $\delta_2 = 0.5(\sqrt{17/6} - \sqrt{1/6})$[20].

2. Compute the averages of the bootstrap responses $\bar{R}_i^b = \frac{1}{n_i}\sum_{i=1}^{n_i} R_{ij}^b$, the bootstrap effect sizes $E_i^b = \bar{R}_i^b - \hat{R}_i^b$ and their standard errors s_i^b, for all off-axis points $i = 1, \ldots, N$.

After the bootstrap procedure has finished, compute the following intervals for a sequence of S constants t^s, $s = 1, \ldots, S$ (e.g. 1, 1.1, 1.15, 1.2, ..., 2.95, 3): $[L_i^b(t^s), U_i^b(t^s)]$, with

$$L_i^b(t) = E_i^b - ts_i^b \quad \text{and} \quad U_i^b(t) = E_i^b + ts_i^b.$$

For each $s = 1, \ldots, S$ and each $b = 1, \ldots, B$, the number of directional false coverages F can be computed as in Equation (4); let F^{sb} denote this number. Averaging over the B bootstrap runs, gives numbers $F^s = \frac{1}{B}\sum_{b=1}^{B} F^{sb}$, and F^s/m may be seen as an approximation of the dFCR if threshold t^s was used for the CI calculations. Now find the smallest t^s that still results in dFCR$\leq \alpha$: $t_\alpha = \min\{t^s : F^s/m \leq \alpha\}$. This is the threshold used for the final calculation of the simultaneous confidence intervals and it will result in the control of the dFCR at the α level. In particular, the CIs are computed as $[L_i, U_i]$ with

$$L_i = E_i - t_\alpha s_i \quad \text{and} \quad U_i = E_i + t_\alpha s_i.$$

Empirical evaluation of the methodology

The formal method for synergy testing, which aims at controlling the dFCR, deviates from what was presented earlier in our papers [12, 17] and hence a thorough evaluation of the new methodology is needed, particularly for an assessment in terms of dFCR control and sensitivity. Since this paper is meant to be a tutorial, we have decided to move the details of the simulation study to supporting information, and report here only the main findings. Briefly, the simulation settings are adopted from the extensive simulation study of our previous work [12].

The results of the simulation studies demonstrate that our procedure succeeds in controlling the dFCR at the nominal level, while showing a sensitivity that is generally larger than what was obtained with our previous testing procedures (maxR and meanR). In the supporting information we give a more detailed discussion on the simulation results.

4 Case study

We will illustrate the described methodology and present data visualizations for synergy analysis using the BIGL R package and data from a drug combination experiment of direct-acting antivirals. Data are from the directAntivirals sample dataset included in the BIGL R package, consisting of 11 drug combination experiments of direct-acting antivirals. To facilitate illustration, we will focus on the 4^{th} experiment. We refer you to the supplementary material for R code implementation.

Measurements from experiments following the designs above often need to be normalized to the control wells. In this 4^{th} experiment, the controls are the wells with 0 dose of both drug A and drug B. We chose to normalize the measurements by taking the ratio of each measurement to the average measurements of the control wells. We will refer to the normalized measurements as responses.

Fitting monotherapies

In the first step, the monotherapy curves for both drugs were estimated utilizing 4PL regression models (see Figure 4). Given the decreasing dose-response curves, a common upper asymptote assumption was made when fitting the model to both drugs. The 4PL regression models fit the data well (responses equally distributed above and below fitted line) and the sigmoidal patterns are well defined (i.e., two points defining upper asymptote, two points defining lower asymptote, and three points defining hill slope) with the selected doses. Notice that the lower asymptotes of Drug A and Drug B are not equal indicating a need for the Generalized Loewe null model.

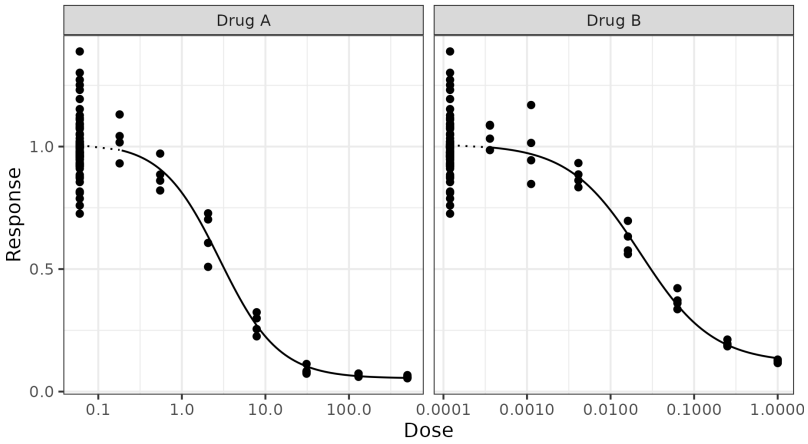


Figure 4: Monotherapy dose-response curves from directAntivirals dataset (BIGL R package), experiment 4. The y-axis is the response (normalized measurement) and the x-axis is the logarithmically transformed dose.

Predicting expected combination response

The expected responses for the chosen null models were calculated at all dose combinations, based on the estimated monotherapy dose-response curves. For illustration purposes, we used three different null models (the HSA, the Bliss, and the Generalized Loewe), enabling a sensitivity analysis of the different assumptions. In Figure 5, the observed and expected responses for the selected null models were visualized with stratifying the response surface by Drug A dose, creating a 2-dimensional trellis plot (the BIGL R package allows either Drug to be used as the stratifying variable). In this experiment, for a particular dose combination, points below the expected response are in the direction of synergy and points above the expected response are in the direction of antagonism. Furthermore, each null model has different underlying assumptions, as such, each null model predicts different expected treatment responses which is clearly visible in trellis 2 (top right facet).

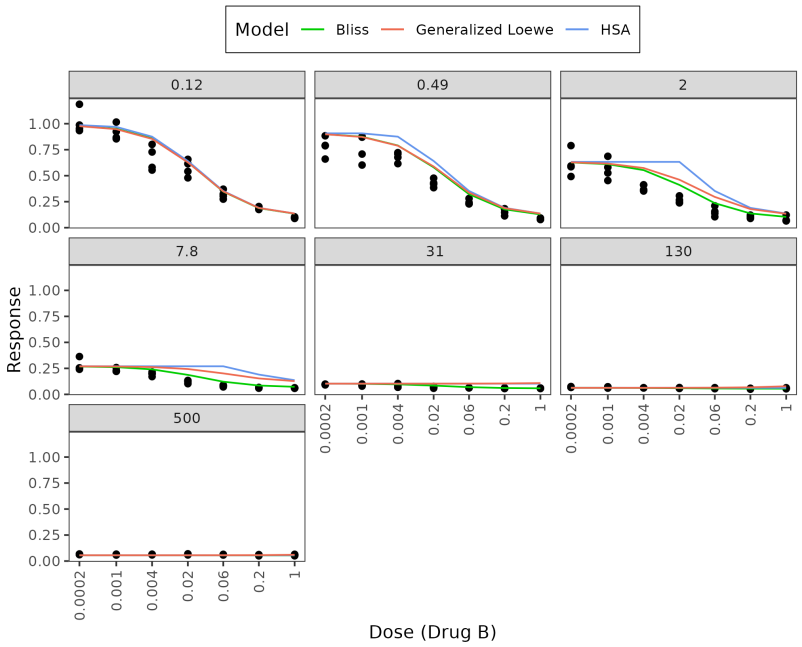


Figure 5: 2-Dimensional stratified predicted response surface plot, stratified by Drug A dose. The points are the observed responses whereas the colored lines are the expected responses derived from the different null models. The panels correspond to dose levels of drug A and the x-axis shows the dose levels of drug B.

Estimating effect sizes

The BIGL R package provides a plot to visualize the effect sizes along with the corresponding confidence intervals for the null models of interest. Simultaneously, the plot allows users to visualize the synergy or antagonism calls by color. Figure 6 represents the effect sizes and corresponding CIs under the Generalized Loewe null model. In particular, a square highlighted in light blue represents a synergy call at the corresponding dose combination. Should antagonism been detected, the corresponding combination would have been highlighted in light pink. For example, synergy was detected at the combination of 2, 0.016 (i.e., dose of Drug A = 2 and dose of Drug B = 0.016). The observed response at this combination was 0.198 units lower than the expected response derived from the Generalized Loewe null model with a 95% confidence interval of (-0.284, -0.111). In addition, synergy/antagonism calls can also be presented in a bi-dimensional contour plot (see Figure 7). The size of the point in the bi-dimensional contour plot represents the magnitude of the effect size. As in Figure 6, the color of the bi-dimensional contour plot indicates the direction of deviance and 0 was not included in the 95% CI. The BIGL R package also allows for these data to be displayed in a table format (supplementary material). All confidence intervals simultaneously control the dFCR at 5%.

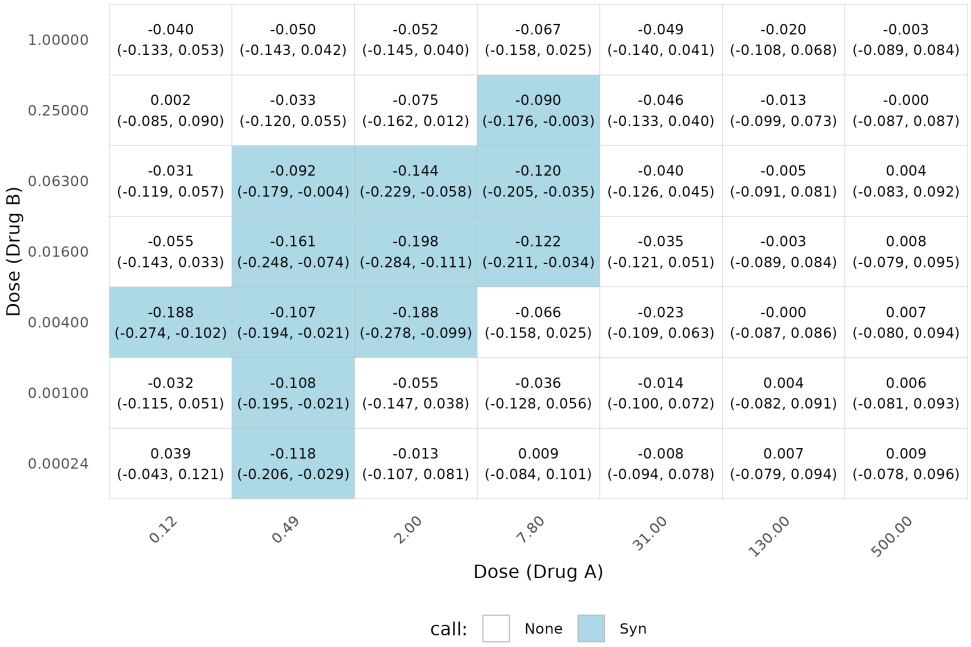


Figure 6: Effect sizes with simultaneous 95% confidence intervals (controlling the dFCR at 5%). Synergy calls are highlighted in light blue.

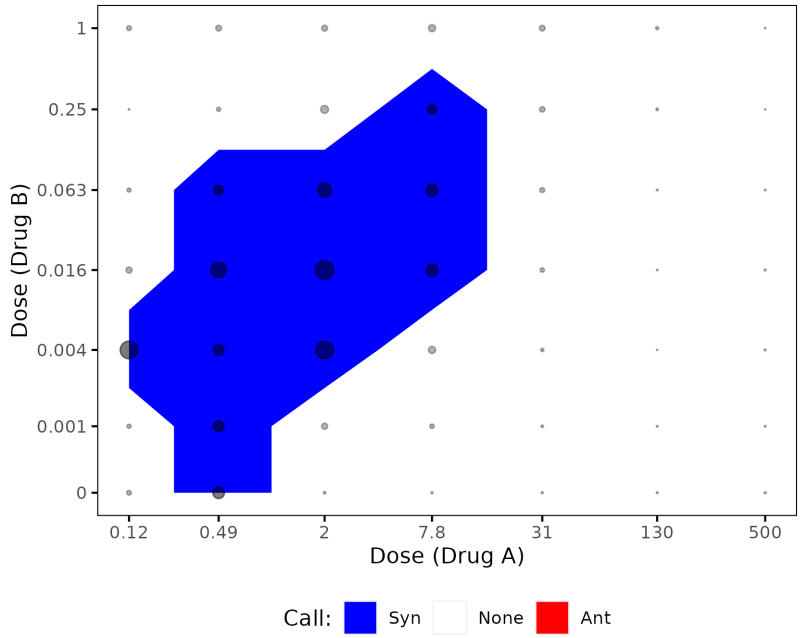


Figure 7: Contour plot. Synergy calls are highlighted in blue. The size of the points are relative to the magnitude of the effect sizes.

5 Discussion

The term synergy is used extensively to justify the potential of drug combinations. However, the meaning of it is often obscured as multiple null models exist, with most frequently used ones described in this tutorial paper. It is important, while evaluating the potential of combinations to keep the assumptions in mind and report accordingly. The HSA null model is the most liberal one, and in a technical sense not truly an evaluation of synergy as it only compares to the maximum response of either one of the monotherapies. Nevertheless, it can be an important evaluation where one would show the benefit of a combination versus a monotherapeutic effect. The other two null models, Bliss and Loewe, evaluate a particular type of additive effect. The Bliss null model is often the most conservative; however, the Bliss null model assumes the two drugs have a different mechanism of action. All these approaches are overly simplistic in that they fail to capture the underlying biological complexity. Hence, if no clear preference on the null model is present based on the assumptions, we suggest to run the models in parallel and evaluate the synergy calls with the assumptions in mind. Often, strong synergistic effects will be detected regardless of null model choice. When synergy is detected only under the HSA null model, this means that the combination only produces an increased effect when compared to a single monotherapy.

As mentioned before, it is important to evaluate the excess response versus the null models considering the variability in the data as observed deviations can be due to inherent variability of the experiment. Hence, our preference goes to methodologies evaluating the variability in the data, and not based on arbitrary thresholds (see Table 1). Additionally, it is important to express the excess response as an interpretable effect size. The latter made us shift from the maxR test in BIGL [17] to reporting the effect sizes together with CIs while controlling the directional false coverage rate (dFCR) as our default methodology. In an in-vitro screening setting, it is important to control the proportion of false calls to avoid validating irrelevant hypothesis downstream while keeping the sensitivity as high as possible. However, in these early screens, it is less crucial to have a precise estimate of the excess itself, as the expected gain of the combination will likely be evaluated downstream in more translatable experiments (e.g. in-vivo). This is why we chose to control the dFCR. In an extensive simulation study, we have demonstrated that our testing procedure controls the dFCR under a wide range of scenarios with acceptable sensitivity. To the best of our knowledge, the evaluation of error rate control, specifically dFCR, which is most important in early stages, is what discriminates BIGL from all other deviance methods described in literature; making it our preferred method.

To maintain a reasonable sensitivity, while controlling the dFCR, it is crucial to minimize the variability in the experiment which can be achieved during assay development and proper experimental design. Firstly, a capture of the full dose response of the monotherapies is required with sufficient data defining the hill slope. Often, the slopes are the areas where synergy could occur as asymptotes generally represent either no or full effect. Synergy calls, in neighbouring sampling points, are more convincing compared to singletons, as biologically, we expect some dynamic concentration ranges where compounds enhance each other. If variability can't be further reduced by identifying sources of variation and controlling key parameters in the assay protocol, it is crucial

to have enough replicates to ensure correct estimation of the variability in the experiment. From experience, we learned that 4 replicates of the full checkerboard is often sufficient. However, research is ongoing to evaluate the required number of replicates that results in adequate sensitivity to detect synergy.

As cancer research is an important field in the exploration of combination therapies and its growing attention in immunotherapies, we feel there is a need to dig deeper in the nature of the replicates. The immune compartment in these type of experiments is often introduced using human donor material. Hence, replicates in this setting, cannot be considered as technical replicates. Replicates derived from biological donors tend to enlarge the variability seen in the experiments. It is crucial to disentangle these different sources of variability. Currently, we are exploring if expectations, based on chosen null models can be calculated within a donor and the excess responses can be pooled between donors for statistical testing.

Finally, the proposed methodology can be extended to three- or four-drug combinations by having one or both, respectively, of the monotherapies be a drug combination. Under this scenario, the EC_{50}s of the monotherapy are less interpretable; however, the methodology remains the same and the effect sizes still represent excess responses.

Acknowledgments

Special thank you to Maxim Nazarov from Open Analytics who is the maintainer of the R-package on CRAN. Additionally, we would like to thank Annelies Tourny and Stijn Hawinkel as they are former developers and their simulation setup remained the foundation of the extensions presented in this paper. Finally, we would like to thank the many scientists using our methodologies for the usefull discussions which encouraged us to continuously improve the methods with clear applications in mind.

Conflicts of Interest

Elli Makariadou, Xuechen Wang, Nicholas Hein and Bie Verbist are currently employed by Janssen Pharmaceutical Companies of Johnson and Johnson. The authors declare no potential conflict of interests.

Data Availability

We refer to the BIGL CRAN R package (`https://cran.r-project.org/web/packages/BIGL/index.html`), which has an embedded directAntivirals dataset including 11 combination experiments.

Supporting information

Simulation Study

We have conducted a simulation study using the same methodology as in our previous work [12]. More specifically, we have implemented scenarios 2 and 3, with slightly different parameter settings: the on-axis standard deviation σ_0 ranges from 0.05 to 0.2 and the numbers of replicates are 2, 3, 4 and 6. The results of the simulation study are also presented in the same way as in [12]: Appendix S1. In this html file the complete set of parameter settings is provided.

The simulation results are evaluated in terms of several criteria. Most of them are the conventional criteria: FDR, sensitivity, specificity, ... However, since we now aim to control the dFCR, we have included a few extra criteria.

The positive predicted value (PPV) and negative predicted value (NPV) refer to the expected proportion of true positive and true negative calls among the positive and negative calls, respectively. More formally, using the notation in Section 3, the PPV is defined as

$$\text{PPV} = \text{E}\left\{\frac{\#\{i : 0 \notin [L_i, U_i] \text{ and } \tau_i \neq 0\}}{\#\{i : 0 \notin [L_i, U_i]\}}\right\}$$

and the NPV is given by

$$\text{NPV} = \text{E}\left\{\frac{\#\{i : 0 \in [L_i, U_i] \text{ and } \tau_i = 0\}}{\#\{i : 0 \in [L_i, U_i]\}}\right\}.$$

We also introduced the concept of *neighbouring* in the evaluation of the method to better reflect a realistic use of the testing procedure. In practice the data analyst often looks at the dose combinations in the (d_1, d_2) plane for which zero is not contained in the CI (i.e. the method gives a synergy or antagonism call). Let us refer to such a point as a *positive point*. If such a positive point is isolated in the sense that no neighbouring points are positive, then many researchers will ignore this point and suspect it as a false positive. A neighbouring point is defined as a point that is within a certain distance in the (d_1, d_2) plane. For example, in a checkerboard design this distance can be chosen such that all closest points along the horizontal, vertical and diagonal directions are considered as neighbouring points. On the other hand, if a positive point has at least one positive neighbouring point, then scientist are willing to believe that these are true deviations from additivity and these points will be considered as *strong positive calls*. In the light of this reasoning we now define the *neighbouring PPV* as

$$\text{nPPV} = \text{E}\left\{\frac{\#\{i : i \in \mathcal{N}^+ \text{ and } \tau_i \neq 0\}}{\#\{i : i \in \mathcal{N}^+\}}\right\},$$

where $\mathcal{N}^+$ is the set of positive off-axis points that have at least one positive neighbour.

In a similar fashion the *neighbouring false discovery rate* (nFDR) is given by

$$\text{nFDR} = \text{E}\left\{\frac{\#\{i : i \in \mathcal{N}^+ \text{ and } \tau_i = 0\}}{\#\{i : i \in \mathcal{N}^+\}}\right\}.$$

Finally, we included criteria related to power. We defined *power3* as the probability to correctly detect at least three synergistic points,

$$\text{power3} = \text{P}\left\{\#\{i : 0 \notin [L_i, U_i] \text{ and } \tau_i \neq 0\} \geq 3\right\}$$

and *power_all* as the probability to detect all synergistic points,

$$\text{power\_all} = \text{P}\left\{\#\{i : 0 \notin [L_i, U_i] \text{ and } \tau_i \neq 0\} = \#\{\tau_i \neq 0\}\right\}.$$

R Code

Load the packages that are needed to generate the results. Set seed to get the same results.

```
library(BIGL)
library(ggplot2)
library(dplyr)
set.seed(1)
```

A function to subset data to a single experiment and, optionally, select the necessary columns only, and create the normalized measurements/responses.

```
subsetData <- function(data, i) {
  subset(data, experiment == i)[, c("effect", "d1", "d2")] %>%
    mutate(effect = effect / mean(effect[d1==0 & d2==0]))
}
```

Extract data of the 4^{th} experiment.

```
data <- subsetData(directAntivirals, 4)
```

Step 1: Fit monotherapy dose-response models

```
mf <- fitMarginals(data, method = "nls", names = c("Drug A", "Drug B"))
```

Step 2: Predict expected combination responses using 3 different null models: the HSA, the Bliss and generalized loewe, controlling the dFCR at 5%.

```
rs <- list()
rs["hsa"]]<-fitSurface(data, mf, null_model="hsa", statistic="both",
                       parallel=4, B.B=20, wild_bootstrap=TRUE,
                       wild_bootType="normal",control="dFCR")
rs["bliss"]]<-fitSurface(data, mf, null_model="bliss", statistic="both",
                         parallel=4, B.B=20, wild_bootstrap=TRUE,
                         wild_bootType="normal",control="dFCR")
rs["loewe"]]<-fitSurface(data, mf, null_model="loewe", statistic="both",
                         parallel=4, B.B=20, wild_bootstrap=TRUE,
                         wild_bootType="normal",control="dFCR")
```

Plot the monotherapy dose-response curves. (Figure 4)

```
plot(mf) + labs(x="Dose", y="Response")
```

Plot the 2-Dimensional stratified predicted response surface plot, stratified by Drug A dose. (Figure 5)

```
synergy_plot_bycomp(rs, color = TRUE, plotBy = "Drug A",
                    xlab = "Dose (Drug B)", ylab="Response")
```

Make the plot of effect sizes with simultaneous 95% confidence intervals. (Figure 6)

```
plotConfInt(rs[["loewe"]], color = "effect-size")
```

Make the contour plot. (Figure 7)

```
contour(rs[["loewe"]], colorBy = "effect-size", digits=3, main = NULL)
```

Print the effect sizes for all the combinations

```
rs[["loewe"]]$confInt$offAxis
```

References

1. Berenbaum MC. What is synergy? Pharmacological reviews 1989;41:93–141.
2. Bliss CI. The toxicity of poisons applied jointly 1. Annals of applied biology 1939;26:585–615.
3. Loewe S. Die quantitativen probleme der pharmakologie. Ergebnisse der Physiologie 1928;27:47–187.
4. Vlot AH, Aniceto N, Menden MP, Ulrich-Merzenich G, and Bender A. Applying synergy metrics to combination screening data: agreements, disagreements and pitfalls. Drug discovery today 2019;24:2286–98.
5. Harbron C. A flexible unified approach to the analysis of pre-clinical combination studies. Statistics in medicine 2010;29:1746–56.
6. Straetemans R, O'Brien T, Wouters L, et al. Design and analysis of drug combination experiments. Biometrical Journal: Journal of Mathematical Methods in Biosciences 2005;47:299–308.
7. Prichard MN and Shipman Jr C. A three-dimensional model to analyze drug-drug interactions. Antiviral research 1990;14:181–205.
8. Ianevski A, He L, Aittokallio T, and Tang J. SynergyFinder: a web application for analyzing drug combination dose–response matrix data. Bioinformatics 2017;33:2413–5.
9. Ianevski A, Giri AK, and Aittokallio T. SynergyFinder 2.0: visual analytics of multi-drug combination synergies. Nucleic acids research 2020;48:W488–W493.
10. Ianevski A, Giri AK, and Aittokallio T. SynergyFinder 3.0: an interactive analysis and consensus interpretation of multi-drug synergies across multiple samples. Nucleic Acids Research 2022;50:W739–W743.

11. Zheng S, Wang W, Aldahdooh J, et al. SynergyFinder plus: toward better interpretation and annotation of drug combination screening datasets. Genomics, Proteomics and Bioinformatics 2022;20:587–96.

12. Thas O, Tourny A, Verbist B, et al. Statistical detection of synergy: New methods and a comparative study. Pharmaceutical Statistics 2022;21:345–60.

13. Di Veroli GY, Fornari C, Wang D, et al. Combenefit: an interactive platform for the analysis and visualization of drug combinations. Bioinformatics 2016;32:2866–8.

14. Turner H, Tourny A, Thas O, et al. BIGL: Biochemically Intuitive Generalized Loewe Model. R package version 1.7.0. 2023. URL: `https://cran.r-project.org/web/packages/BIGL/index.html`.

15. Baud M. Methods of Immunological Analysis 1: Fundamentals. 1993.

16. Sebaugh J. Guidelines for accurate EC50/IC50 estimation. Pharmaceutical statistics 2011;10:128–34.

17. Van der Borght K, Tourny A, Bagdziunas R, et al. BIGL: Biochemically Intuitive Generalized Loewe null model for prediction of the expected combined effect compatible with partial agonism and antagonism. Scientific Reports 2017;7:17935.

18. Benjamini Y and Yekutieli D. False discovery rate–adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association 2005;100:71–81.

19. Benjamini Y, Hochberg Y, and Kling Y. False discovery rate control in pairwise comparisons. Research Paper, Department of Statistics and Operations Research, Tel Aviv University 1993.

20. Liu RY. Bootstrap procedures under some non-iid models. The annals of statistics 1988;16:1696–708.