pharmaceuticals

MDPI

Article

The Role of Outcome Response Rate in Planning Biosimilar Studies Using Different Evaluation Metrics

Liyi Cen *, Ramin Arani and Dejun Tang

Sandoz Inc., Princeton, NJ 08540, USA
* Correspondence: liyi.cen@sandoz.com

Abstract: Background/Objectives: Biosimilar studies use overall response rate to assess clinical similarity. Sample size and power depend on the equivalence margin, defined in either risk difference or risk ratio scale. This manuscript investigates how different evaluation metrics and varying response rates affect study power. **Methods:** Two numerical simulations are conducted. The first is designed to test in the risk difference scale, while the second tests in the risk ratio scale. Both simulations consider no difference between the biosimilar and reference product. Response rates vary from 0.1 to 0.9, and all scenarios are repeated 10,000 times. **Results:** The study shows inconsistent results in testing the equivalence of overall response rate across the risk difference and risk ratio scales, even when the hypotheses are mathematically equivalent. Consequently, the study is often under powered for testing in both scales. Additionally, study power is sensitive to outcome response rate deviation, with different directions of change in the two different evaluation metrics. **Conclusions:** Biosimilar study design should avoid the concept of converting equivalence margins between risk difference and risk ratio scales, assuming no change in study power. Careful strategies should be planned for estimating overall response rates for sample size assessments.

Keywords: bioequivalence study; risk difference; risk ratio; equivalence margin conversion; overall response rate; biosimilar

check for updates

Academic Editor: Giovanni M. Pitari

Received: 1 January 2025
Revised: 7 February 2025
Accepted: 10 February 2025
Published: 12 February 2025

Citation: Cen, L.; Arani, R.; Tang, D. The Role of Outcome Response Rate in Planning Biosimilar Studies Using Different Evaluation Metrics. *Pharmaceuticals* **2025**, *18*, 243. https://doi.org/10.3390/ph18020243

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In immuno-oncology biosimilar studies, the commonly used primary endpoint is the overall response rate within a specific time period to assess clinical similarity between a biosimilar and its reference product [1,2]. To demonstrate efficacy similarity, it is required to demonstrate that the clinical effect endpoints, for instance, the overall response rate, are equivalent for the proposed biosimilar and its reference product, measured by their corresponding predefined margins [1,3–6]. When the primary efficacy endpoint is chosen, a crucial aspect of designing a biosimilar study is to estimate the treatment effect and determine the equivalence margin.

The equivalence margin is closely related to the treatment effect size of the reference product. During the study design, the treatment effect size for the reference product is usually obtained by the lower limit of its 95% confidence interval, which is typically calculated through a meta-analysis of available data from historical studies [1,3]. Subsequently, the equivalence margin is set to preserve a certain percentage (typically 50%) of that effect size [3,4]. Based on the equivalence margin and the assumed treatment effect, the sample size and the power of the study can then be determined.

When using the overall response rate as the key efficacy endpoint, two commonly used evaluation metrics for comparing it between a biosimilar and its reference product are absolute risk difference and relative risk difference (or risk ratio) [7–10]. However, the choice of equivalence margin in these two metrics is subject to controversy, with no consensus reached within the scientific community or regulatory agencies. The scientific community has called for consensus on this matter [11]. Nevertheless, it is still common to design studies using both metrics to comply with requirements from various regulatory agencies, for example, in the several trastuzumab biosimilar clinical developments [9,12–21].

Once an equivalence margin is set for a trial, it is crucial to ensure the study has sufficient power to claim that the proposed biosimilar treatment is not different from the reference product. However, striking a balance between study power and a feasible sample size is a challenging decision for sponsors, especially for immuno-oncology studies [9,12]. As a key factor for the study power, outcome response rates are typically estimated based on historical data during the study design, but actual rates may differ from what are assumed in sample size assessments [7]. Consequently, the study power may vary depending on the actual response rate. The impact of this variation on study power is unclear, and it is also uncertain whether the same impact would occur if the statistical tests were conducted on risk difference or risk ratio scales.

There are two objectives for this manuscript. Both are related to the role of outcome response rate in designing a bioequivalence trial with two different evaluation metrics: risk difference and risk ratio. The first objective is to investigate the study power when the hypothesis testing is performed in the risk difference and risk ratio scale, respectively. We will examine the concordance of the two tests in different settings and explore the role of the overall response rate in this relationship. Second, we will investigate how the study power changes when the actual response rate differs from the assumed one during study design.

2. Results

2.1. Discordance of Equivalence Test in Risk Difference and Risk Ratio Scales

Table 1 presents the test results obtained in both risk difference and risk ratio scales when the study was originally designed to assess the risk difference. The equivalence margin for the risk ratio scale is derived from the equivalence margin of $[-0.05, 0.05]$ in the risk difference scale, as outlined in Section 4.2. It is noted that the equivalence margin in the risk ratio scale varies for different response rates and is no longer symmetric. For each response rate, we analyzed the data in both risk difference and risk ratio scales and repeated the study 10,000 times. Table 1 reports the sample size per arm, converted equivalence margin in the risk ratio scale and the percentage of time that the study yields a positive result (i.e., null hypothesis is rejected) for the two tests in risk difference and risk ratio scales. Additionally, it reports the concordance between the two tests. The percentage of positive results for the test in the risk difference scale is very close to 80%, which aligns with the designed study power of 80%. Similarly, the percentage of positive results for the test in the risk ratio scale is also close to 80%, although it tends to be slightly lower compared to the test in the risk difference scale, particularly when the response rate is low. The discordance between the two tests ranges from 3.4% to 24.9%. A lower response rate is associated with a higher discordance. Consequently, the percentage of positive results for both tests is significantly lower than 80% when the response rate is low.

Table 1. Comparison of equivalence test results in risk difference and risk ratio scales in study design I. The sample size is selected such that the study power is 80% when testing in the risk difference scale with type I error rate of 0.05. All scenarios were repeated 10,000 times.

Response Rate	Sample Size Per Arm	RR Margin [1,2]		Percent Time with Positive Test Result					
				RD Test	RR Test	Both Test	RD Test Only	RR Test Only	Discordance [3]
0.1	757	0.50	1.50	79.7%	73.7%	64.3%	15.5%	9.4%	24.9%
0.2	1345	0.75	1.25	79.8%	78.1%	72.1%	7.8%	6.1%	13.8%
0.3	1766	0.83	1.17	79.4%	79.1%	74.6%	4.8%	4.5%	9.3%
0.4	2018	0.88	1.13	79.7%	79.1%	75.9%	3.8%	3.2%	7.0%
0.5	2102	0.90	1.10	80.1%	79.8%	77.1%	3.0%	2.6%	5.6%
0.6	2018	0.92	1.08	80.0%	79.8%	77.5%	2.6%	2.3%	4.8%
0.7	1766	0.93	1.07	79.9%	79.8%	78.0%	1.9%	1.8%	3.6%
0.8	1345	0.94	1.06	80.2%	80.2%	78.4%	1.8%	1.8%	3.6%
0.9	757	0.94	1.06	80.0%	79.7%	78.1%	1.8%	1.6%	3.4%

[1] RD: risk difference; RR: risk ratio. [2] The equivalence margin for risk ratio is calculated from the equivalence margin for risk difference using the formula in Section 4.2. [3] The discordance is the sum of the "RD test only" and "RR test only" columns.

Table 2 presents the test results obtained in both risk difference and risk ratio scales when the study was originally designed to assess the risk ratio. The equivalence margin for the risk difference scale is derived from the equivalence margin of $[\frac{1}{1.2}, 1.2]$ in the risk ratio scale, as outlined in Section 4.2. It also varies for different response rates and is no longer symmetric. The percentage of positive results for the test in the risk ratio scale is very close to 80%, which aligns with the designed study power of 80%. Similarly, the percentage of positive results for the test in the risk difference scale is also close to 80%. The discordance between the two tests is around 10% in all scenarios. Consequently, the percentage of positive results for both tests is a few percent lower than 80%.

Table 2. Comparison of equivalence test results in risk difference and risk ratio scales in study design II. The sample size is selected such that the study power is 80% when testing in the risk ratio scale with type I error rate of 0.05. All scenarios were repeated 10,000 times.

Overall Response Rate	Sample Size Per Arm	RD Margin [1,2]		Percent Time with Positive Test Result					
				RD Test	RR Test	Both Test	RD Test Only	RR Test Only	Discordance [3]
0.1	5712	−0.02	0.02	79.5%	80.5%	74.7%	4.7%	5.8%	10.5%
0.2	2530	−0.03	0.04	78.3%	79.7%	74.0%	4.4%	5.8%	10.1%
0.3	1483	−0.05	0.06	79.1%	79.8%	74.3%	4.9%	5.5%	10.4%
0.4	954	−0.07	0.08	79.1%	80.3%	74.6%	4.5%	5.7%	10.2%
0.5	628	−0.08	0.10	78.8%	79.6%	73.8%	5.0%	5.8%	10.9%
0.6	418	−0.10	0.12	79.0%	79.4%	73.9%	5.1%	5.5%	10.6%
0.7	274	−0.12	0.14	79.0%	79.9%	74.4%	4.6%	5.5%	10.1%
0.8	160	−0.13	0.16	80.2%	80.0%	75.0%	5.2%	5.0%	10.2%
0.9	72	−0.15	0.18	80.7%	80.7%	75.7%	5.0%	5.0%	10.0%

[1] RR: risk difference; RD: risk ratio. [2] The equivalence margin for risk difference is calculated from the equivalence margin for risk ratio using the formula in Section 4.2. [3] The discordance is the sum of the "RD test only" and "RR test only" columns.

Figure 1 illustrates the comparison between the outcome response rate and the percentage of positive results for both tests. When the study is initially designed to test in the risk ratio scale (Figure 1, right panel), the percentage of positive results remains relatively consistent around 75%. However, when the study is initially designed to test in the risk difference scale (Figure 1, left panel), the percentage of positive results varies significantly with the response rate. Notably, a significantly low percentage of positive results for both tests is observed when the response rate is low in the latter case.

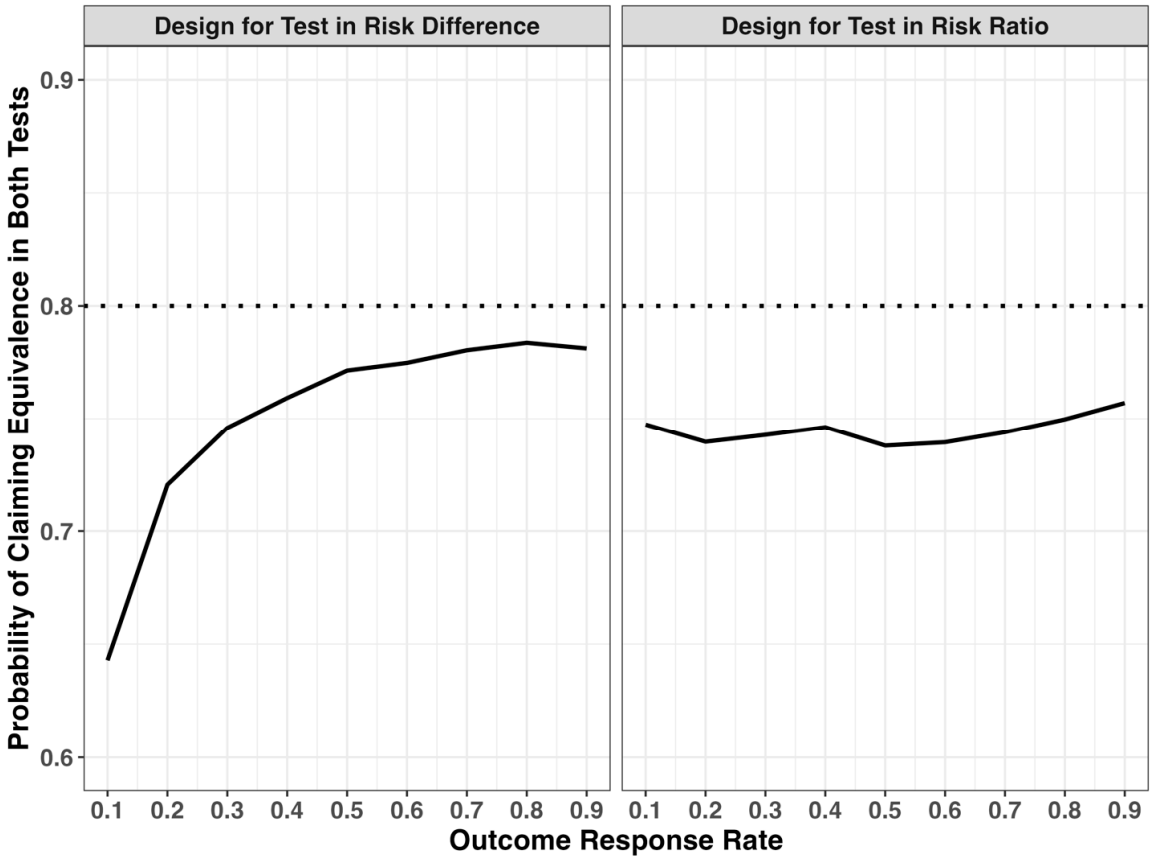

Figure 1. Percentage of positive results for both tests in risk difference and risk ratio scales when (**left**) the study is originally designed to test in risk difference and (**right**) the study is originally designed to test in risk ratio.

2.2. Sensitivity of Study Power When the Outcome Response Rate Deviates from the Assumed Level During Study Design

When the study is initially designed to test for the risk difference, the study power with respect to the assumed outcome response rate during the study design is illustrated at different levels of deviations in Figure 2. If the outcome response rate is overestimated, meaning the actual outcome response rate is lower than the originally assumed rate (e.g., the red line in Figure 2 represents the scenario in which the actual outcome response rate is 0.05 lower than the assumed level), the study power is larger than the planned 80% level when the outcome response rate is below 0.5. Conversely, the study power is smaller than the planned 80% level when the outcome response rate is above 0.5. If the outcome response rate is underestimated, meaning the actual outcome response rate is higher than the assumed rate (e.g., the purple line in Figure 2 represents the scenario in which the actual outcome response rate is 0.05 higher than the assumed level), the study power is smaller than the planned 80% level when the outcome response rate is less than 0.5. The study power is larger than the planned 80% level if the outcome response rate is greater than 0.5. Furthermore, the study

power remains relatively stable regardless of the change from the assumed response rate if the outcome response rate is close to 0.5. For example, with the assumed response rate of 0.5, the study power is 81% or 80.4% if the response rate reduces to 0.45 or increases to 0.55, respectively (Table 3, Study Design I). However, if the outcome response rate is either low (close to 0) or high (close to 1), the study power can change quite a bit, even for a small change in the response rate. For example, with the assumed response rate of 0.2, the study power increases to 90.2% if the response rate reduces to 0.15 (a change of −0.05). The study power reduces to only 69.9% if the response rate increases to 0.25 (a change of 0.05).

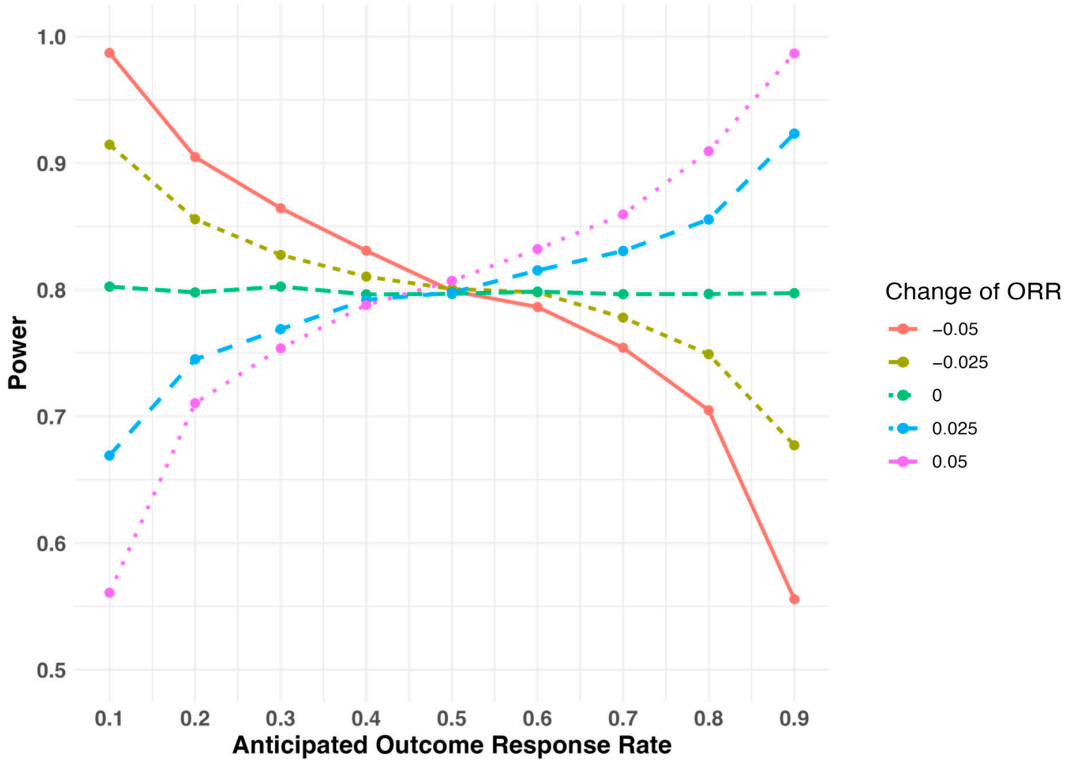

Figure 2. Study power with respect to the assumed outcome response rate (ORR) at different levels of deviations when the study is initially designed to test in the risk difference scale.

Table 3. Power of study when the observed response rate is deviated from the expected response rate by 0.05.

Assumed Response Rate	Study Design I		Study Design II	
	Actual Response Rate Is 0.05 Smaller Than Expected	Actual Response Rate Is 0.05 Larger Than Expected	Actual Response Rate Is 0.05 Smaller Than Expected	Actual Response Rate Is 0.05 Larger Than Expected
0.1	98.8%	55.4%	22.3%	96.5%
0.2	90.2%	69.9%	56.1%	92.4%
0.3	85.8%	75.5%	62.6%	90.3%
0.4	83.8%	77.9%	66.4%	90.0%
0.5	81.0%	80.4%	66.1%	89.4%
0.6	78.7%	83.2%	66.4%	89.3%
0.7	75.1%	86.1%	64.7%	91.0%
0.8	70.5%	90.8%	61.2%	93.3%
0.9	55.6%	98.6%	48.7%	98.5%

When the study is designed to test for the risk ratio, the relationship between study power and the change in outcome response rate is quite different (Figure 3). The study power increases with an increase in the outcome response rate and decreases with a decrease in the outcome response rate, regardless of its absolute level. Moreover, the change in study power is larger when the outcome response rate is either low (close to 0) or high (close to 1) compared to when it is close to 0.5. For example, with the assumed response rate of 0.5, the study power reduces to 66.1% if the response rate reduces to 0.45 by 0.05. With the assumed response rate of 0.2, the study power reduces to 56.1% if the response rate reduces to 0.15 by 0.05 (Table 3, Study Design II).

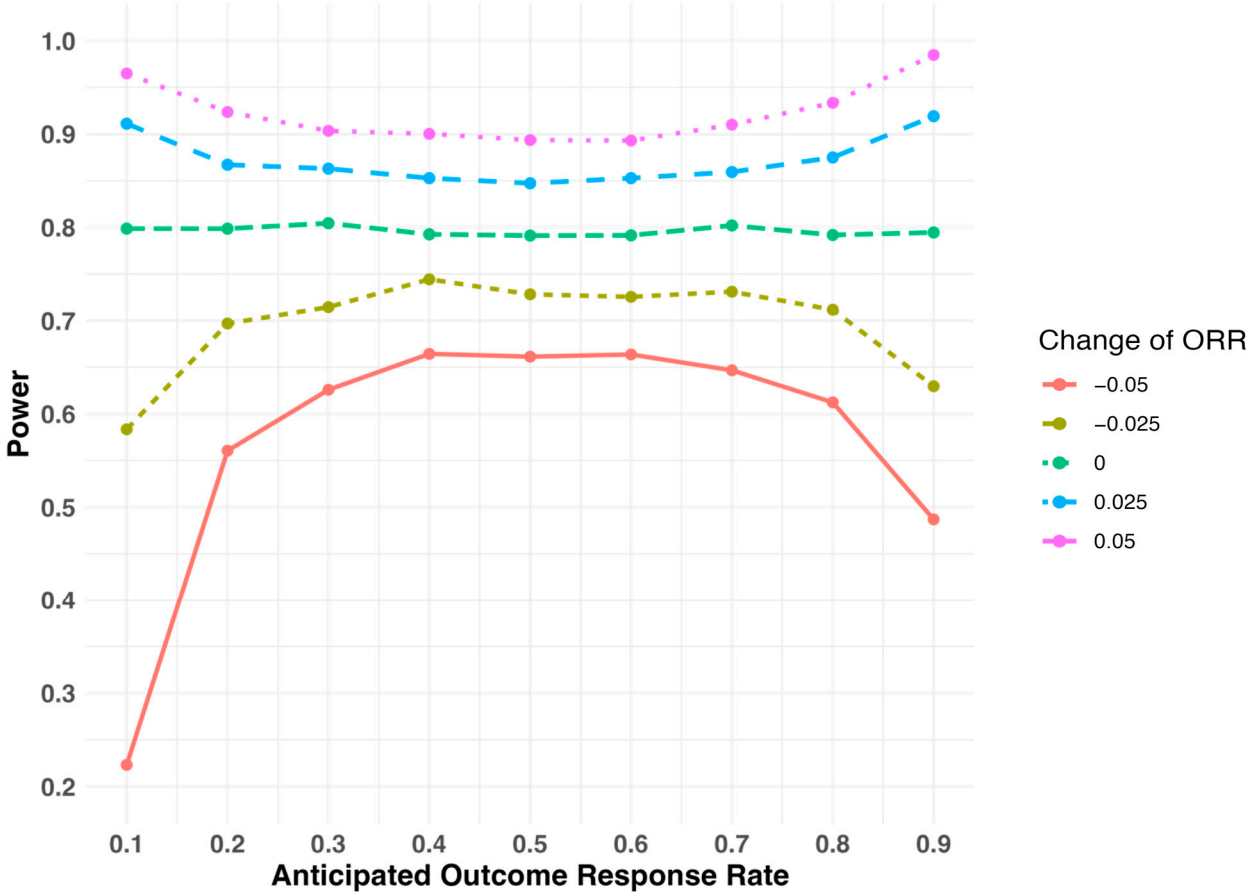

Figure 3. Study power with respect to the assumed outcome response rate (ORR) at different levels of deviations when the study is initially designed to test in the risk ratio scale.

3. Discussions

In this manuscript, we demonstrate that the study lacks sufficient power to detect equivalence using both metrics if it was initially designed to test only in one metric due to the discordance of the two test results. Additionally, the study power can be sensitive to the outcome response rate when it deviates from the assumed level. Furthermore, the direction of change in power can differ depending on whether the risk difference or risk ratio metric is used.

We employed two study designs to compare the study power evaluated in the risk difference and risk ratio scales as well as the concordance of the test results between the two. When using the equivalence margins between the two metrics that maintain the same hypotheses, we discovered significant discordance between the two tests. The magnitude of this discordance depends on the outcome response rate and the original evaluation metric used in designing the study. Consequently, if a study was originally designed to have

sufficient power in one test only, it may be severely underpowered to claim equivalence in both tests due to this discordance. This finding has significant implications for designing the study. It is crucial to recognize that, if a study concludes equivalence using one metric, it does not necessarily imply the same conclusion using the other metric, even if the equivalence margin is calculated to ensure the same hypotheses are tested between the two metrics. Therefore, when designing a biosimilar study, it is essential to avoid the concept of conversion in the equivalence margin and assume that the study power will remain the same.

Instead, if the sample size is calculated to ensure sufficient power for claiming equivalence in one test, it must be increased to achieve the same level of power for both tests. The extent of this sample size increase depends on which test the study was originally powered for and, possibly, the outcome response rate. If we begin with a study design that ensures sufficient power for testing in the risk ratio scale, the percentage increase in sample size remains relatively constant regardless of the outcome response rate. In our numerical simulations, we found that a 10% increase in sample size ensures the study has at least 80% power to claim equivalence in both tests. However, if the study was originally designed to ensure sufficient power for testing in the risk difference scale, the percentage increase in sample size also depends on the outcome response rate. A lower outcome response rate necessitates a larger sample size increase. In practical terms, it may be more operationally convenient to calculate the sample size in the risk ratio scale and then increase it by a specific percentage to ensure sufficient power in both tests, assuming that the equivalence margin defined by the mathematically converted upper and lower bound in the risk difference scale is clinically acceptable.

Sponsors may also consider the impact on the study duration and cost with the increase in sample size. Moore et al. [22] reported that the median enrollment for 24 biosimilar phase 3 trials conducted for products approved between January 2010 and October 2019 was 538 participants. The median trial completion time of these studies was 26 months, with a median cost of $27.6 million. Assuming a proportional relationship between the sample size, cost and time, a 10% increase in sample size would lead to an estimated cost increase of $2.76 million and extend the trial duration by roughly 2.6 months. However, it is important to note that the actual impact may vary depending on different factors, such as therapeutic indication, recruitment rate and operational cost.

The simulation results presented in this manuscript are from the parallel design. The same principle would apply if other types of designs (e.g., crossover or hybrid parallel-crossover design) need to use both metrics to evaluate the similarity. It is recommended to consider the equivalence margin carefully in different metrics instead of simply converting the margin from one scale into the other.

During clinical study design, it is common to assume a fixed outcome response rate to determine the sample size. However, the actual response rate can vary. Macaya et al. [11] performed a literature review from 2010 to 2015 and identified nine non-inferiority trials comparing new-generation to second-generation stents. They reported that the observed event rate was lower than expected in all but one study. In some trials, the difference was substantial. As a result, only four of the nine trials consistently demonstrated non-inferiority using the relative risk metric as compared to the original margin defined in rate difference in those studies. The two numerical studies in this manuscript illustrated how the response rate affects study power when using different evaluation metrics. When the outcome response rate is below 0.5, the study power calculated using equivalence margins defined in the two metrics changes in opposite directions with the response rate. An increase in the response rate leads to reduced power for the test in risk difference and increased power for the test in risk ratio. In this case, a more precise estimate of the

outcome response rate is crucial to maintain the study power as intended. On the other hand, when the outcome response rate is above 0.5, the study power for both metrics changes in the same direction with the change in outcome response rate. An increase in the response rate leads to an increase in the study power. In this situation, it is better to be conservative, assuming a lower response rate for sample size assessments. Li et al. [7] suggested that different scenarios of event rates should be considered at the design stage to ensure adequate power for the chosen margin in non-inferiority trials. Our finding is also consistent with the recommendation for equivalence trials, particularly for trials using both RR and RD metrics as required by different health authorities.

4. Methods

4.1. Equivalence Test of Outcome Response Rate

We first review how the hypothesis test is performed using the risk difference metric. Let p_t and p_r denote the outcome response rate in the biosimilar arm and the reference arm, respectively. When the risk difference evaluation metric is used, the null (H_0) and alternative (H_1) hypotheses for equivalence studies are as follows:

$$H_0 : \ p_t - p_r \leq \delta_L \text{ or } p_t - p_r \geq \delta_U$$

$$H_1 : \ \delta_L < p_t - p_r < \delta_U$$

where $[\delta_L, \ \delta_U]$ is a pre-specified equivalence margin. In biosimilar development, the target risk difference is zero in most cases so that the equivalence margin is often symmetric around zero, i.e., $\delta_U = -\delta_L$. If we set $\delta_U = \delta$, the equivalence margin is $[-\delta, \delta]$.

The equivalence test is usually performed using two one-sided tests [23]. For the left-sided test, the null and alternative hypotheses are

$$H_{0,L} : \ p_t - p_r \leq -\delta \text{ vs. } H_{1,L} : \ p_t - p_r > -\delta$$

For the right-sided test, the null and alternative hypotheses are

$$H_{0,U} : \ p_t - p_r \geq \delta \text{ vs. } H_{1,U} : \ p_t - p_r < \delta$$

The equivalence test can be carried out similarly using the risk ratio evaluation metric. The null and alternative hypotheses are

$$H_0 : \ \frac{p_t}{p_r} \leq \lambda_L \text{ or } \frac{p_t}{p_r} \geq \lambda_U$$

$$H_1 : \ \lambda_L < \frac{p_t}{p_r} < \lambda_U$$

where $[\lambda_L, \ \lambda_U]$ is the equivalence margin defined in the risk ratio scale. The target risk ratio is usually 1 so that, for a symmetric equivalence margin, we can set $\lambda_L = \frac{1}{\lambda_U}$ (assume that $\lambda_U > 1$). If we set $\lambda_U = \lambda$, the symmetric equivalence margin becomes $[\frac{1}{\lambda}, \lambda]$.

Similar to the test in risk difference scale, the test in risk ratio scale can also be carried out using two one-sided tests. For the left-sided test, the null and alternative hypotheses are

$$H_{0,L} : \ \frac{p_t}{p_r} \leq \frac{1}{\lambda} \text{ vs. } H_{1,L} : \ \frac{p_t}{p_r} > \frac{1}{\lambda}$$

For the right-sided test, the null and alternative hypotheses are

$$H_{0,U} : \ \frac{p_t}{p_r} \geq \lambda \text{ vs. } H_{1,U} : \ \frac{p_t}{p_r} < \lambda$$

4.2. Equivalence Margin Conversion Between Risk Difference and Risk Ratio Scale

Biosimilar studies sometimes aim to establish equivalence in both risk difference and risk ratio scales to meet requirements from different regulatory agencies. This would require the specification of equivalence margin for the risk difference and risk ratio separately in one study. Under certain conditions, the null and alternative hypotheses in the two one-sided tests in the risk difference scale are equivalent to the two one-sided tests in the risk ratio scale as follows.

Assume the overall response rate for the reference product is p, and we design the study based on the risk difference and pre-specify $[-\delta,\ \delta]$ as the equivalence margin for the test in risk difference scale. The alternative hypothesis states that the overall response rate for the biosimilar product falls within $[p-\delta,\ p+\delta]$, which is equivalent to saying that the risk ratio between the two products falls within $[\frac{p-\delta}{p},\ \frac{p+\delta}{p}]$. Therefore, if we set the equivalence margin for the risk ratio as $[1-\frac{\delta}{p},\ 1+\frac{\delta}{p}]$, the two hypothesis tests in the risk difference and risk ratio scale are equivalent. However, note that the equivalence margin for the risk ratio is not symmetric anymore. Although the hypothesis tests are mathematically equivalent in the two scales, the power of the equivalent test for the risk ratio will be different since the margin is asymmetric.

Similarly, we can perform the test using the risk ratio metric with a pre-specified equivalence margin for the risk ratio as $[\frac{1}{\lambda},\lambda]$ (assume $\lambda > 1$). The alternative hypothesis states that the overall response rate for the biosimilar product falls within $[\frac{p}{\lambda},\ p\lambda]$, which is equivalent to saying that the risk difference between the two products falls within $[\frac{p}{\lambda}-p,\ p\lambda-p]$. Therefore, if we set the equivalence margin for the risk difference as $[\left(\frac{1}{\lambda}-1\right)p, (\lambda-1)p]$, the two hypothesis tests in the risk difference and risk ratio scale are again equivalent. Again, the equivalence margin for the risk difference is not symmetric anymore. Due to the same reason mentioned above, the power of the test for the risk difference will be different due to the asymmetric margin.

We will use this as our basis for relating the equivalence margin from one scale to the other in the numerical simulation studies.

4.3. Simulation Study Setup

We conduct two simulation studies to investigate the two study objectives. Table 4 outlines the design of the two studies. The first study is designed to perform an equivalence test using the risk difference metric, while the second study is designed to perform the test using the risk ratio metric. We consider a series of response rates for the reference product between 0.1 and 0.9. We consider the case when there is no difference in the overall response rate between the biosimilar and reference product (i.e., the alternative hypothesis is true).

Table 4. Summary of study design.

	Study Design I	**Study Design II**
Metric of evaluation	Risk Difference	Risk Ratio
Reference response rate	10% to 90%	10% to 90%
Expected difference between treatments	RD = 0	RR = 1
Equivalence margin	$[-0.05, 0.05]$	$[\frac{1}{1.2}, 1.2]$
Type I error	0.05	0.05
Study power	0.8	0.8

For the first study, we set the equivalence margin as $[-0.05, 0.05]$ on the risk difference scale. For the second study, we set the equivalence margin as $[\frac{1}{1.2}, 1.2]$ on the risk ratio scale. Based on the assumed response rates and equivalence margins, we calculate the sample size to ensure that the test has 80% power in the originally selected scale (risk difference scale in study I and risk ratio scale in study II) with a type I error rate of 0.05. To assess the study power in scales other than the original design, we use the equivalence margin defined by the mathematically converted lower and upper bound suggested in the previous section.

Here is the summary of simulation steps: (1) simulate binary outcomes with predetermined response rates and sample sizes for both treatment arms; (2) estimate the risk difference, risk ratio, and corresponding 95% confidence intervals from the simulated data; (3) compare the 95% confidence interval with the equivalence margin to determine if the equivalence is established in the risk difference and risk ratio scale; (4) compare the conclusions from the two tests conducted in the two different scales. The R program used for the simulation is provided in the Supplementary Materials.

Additionally, we evaluate the study power when the observed response rate differs from the assumed response rate during study design. We conduct two tests in both the risk difference and risk ratio scales, varying the response rate deviation among four levels: -0.05, -0.025, 0.025 and 0.05. In both studies, all scenarios are repeated 10,000 times.

5. Conclusions

In summary, when designing a biosimilar study, it is essential to avoid the concept of conversion in the equivalence margin between the risk difference and risk ratio scales and assume that the study power remains the same. Additionally, a careful strategy should be considered for estimating the overall response rate for the purpose of sample size assessments.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ph18020243/s1. The supplementary section includes the R program used for the simulation.

Author Contributions: Conceptualization, L.C. and D.T.; Methodology, L.C., R.A. and D.T.; Formal analysis, L.C.; Writing—original draft, L.C.; Writing—review & editing, R.A. and D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article or Supplementary Materials.

Conflicts of Interest: L.C. and R.A. are employed by Sandoz Inc. The work was conducted while D.T. was employed by Sandoz Inc.

References

1. U.S. Food and Drug Administration. Scientific Considerations in Demonstrating Biosimilarity to a Reference Product. 2015. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/scientific-considerations-demonstrating-biosimilarity-reference-product (accessed on 20 December 2024).
2. Uno, H.; Schrag, D.; Kim, D.H.; Tang, D.; Tian, L.; Rugo, H.S.; Wei, L.-J. Assessing Clinical Equivalence in Oncology Biosimilar Trials with Time-to-Event Outcomes. *JNCI Cancer Spectr.* **2019**, *3*, pkz058. [CrossRef] [PubMed]
3. European Medicines Agency. Biosimilars in the EU: Information Guide for Healthcare Professionals. 2017. Available online: https://www.ema.europa.eu/en/documents/leaflet/biosimilars-eu-information-guide-healthcare-professionals_en.pdf (accessed on 20 December 2024).

4. U.S. Food and Drug Administration. Non-Inferiority Clinical Trials to Establish Effectiveness: Guidance for Industry. 2016. Available online: https://www.fda.gov/media/78504/download (accessed on 20 December 2024).
5. U.S. Food and Drug Administration. Statistical Principles for Clinical Trials. 1998. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials (accessed on 20 December 2024).
6. Walker, E.; Nowacki, A.S. Understanding Equivalence and Noninferiority Testing. *J. Gen. Intern. Med.* **2011**, *26*, 192–196. [CrossRef]
7. Li, Z.; Quartagno, M.; Böhringer, S.; van Geloven, N. Choosing and Changing the Analysis Scale in Non-Inferiority Trials with a Binary Outcome. *Clin. Trials* **2022**, *19*, 14–21. [CrossRef] [PubMed]
8. Wellek, S. Statistical Methods for the Analysis of Two-Arm Non-Inferiority Trials with Binary Outcomes. *Biom. J. J. Math. Methods Biosci.* **2005**, *47*, 48–61. [CrossRef] [PubMed]
9. Isakov, L.; Jin, B.; Jacobs, I.A. Statistical Primer on Biosimilar Clinical Development. *Am. J. Ther.* **2016**, *23*, e1903–e1910. [CrossRef] [PubMed]
10. Hilton, J.F. Noninferiority Trial Designs for Odds Ratios and Risk Differences. *Stat. Med.* **2010**, *29*, 982–993. [CrossRef]
11. Macaya, F.; Ryan, N.; Salinas, P.; Pocock, S.J. Challenges in the Design and Interpretation of Noninferiority Trials: Insights from Recent Stent Trials. *J. Am. Coll. Cardiol.* **2017**, *70*, 894–903. [CrossRef] [PubMed]
12. Dettori, J.R.; Norvell, D.C.; Chapman, J.R. Is the Sample Size Big Enough? 4 Things You Need to Know! *Glob. Spine J.* **2022**, *12*, 1027–1028. [CrossRef]
13. Barbier, L.; Declerck, P.; Simoens, S.; Neven, P.; Vulto, A.G.; Huys, I. The Arrival of Biosimilar Monoclonal Antibodies in Oncology: Clinical Studies for Trastuzumab Biosimilars. *Br. J. Cancer* **2019**, *121*, 199–210. [CrossRef] [PubMed]
14. Pivot, X.; Bondarenko, I.; Nowecki, Z.; Dvorkin, M.; Trishkina, E.; Ahn, J.-H.; Vinnyk, Y.; Im, S.-A.; Sarosiek, T.; Chatterjee, S.; et al. Phase III, Randomized, Double-Blind Study Comparing the Efficacy, Safety, and Immunogenicity of SB3 (Trastuzumab Biosimilar) and Reference Trastuzumab in Patients Treated with Neoadjuvant Therapy for Human Epidermal Growth Factor Receptor 2–Positive Early Breast Cancer. *J. Clin. Oncol.* **2018**, *36*, 968–974. [PubMed]
15. Von Minckwitz, G.; Ponomarova, O.; Morales, S.; Zhang, N.; Hanes, V. Efficacy and Safety of Biosimilar ABP 980 Compared with Trastuzumab in HER2 Positive Early Breast Cancer. *Ann. Oncol.* **2017**, *28*, v44. [CrossRef]
16. von Minckwitz, G.; Colleoni, M.; Kolberg, H.-C.; Morales, S.; Santi, P.; Tomasevic, Z.; Zhang, N.; Hanes, V. Efficacy and Safety of ABP 980 Compared with Reference Trastuzumab in Women with HER2-Positive Early Breast Cancer (LILAC Study): A Randomised, Double-Blind, Phase 3 Trial. *Lancet Oncol.* **2018**, *19*, 987–998. [CrossRef]
17. Rugo, H.S.; Barve, A.; Waller, C.F.; Hernandez-Bronchud, M.; Herson, J.; Yuan, J.; Manikhas, A.; Bondarenko, I.; Mukhametshina, G.; Nemsadze, G.; et al. Heritage: A Phase III Safety and Efficacy Trial of the Proposed Trastuzumab Biosimilar Myl-1401O versus Herceptin. *J. Clin. Oncol.* **2016**. Available online: https://ascopubs.org/doi/10.1200/JCO.2016.34.18_suppl.LBA503 (accessed on 30 December 2024). [CrossRef]
18. Rugo, H.S.; Barve, A.; Waller, C.F.; Hernandez-Bronchud, M.; Herson, J.; Yuan, J.; Sharma, R.; Baczkowski, M.; Kothekar, M.; Loganathan, S.; et al. Effect of a Proposed Trastuzumab Biosimilar Compared with Trastuzumab on Overall Response Rate in Patients with ERBB2 (HER2)–Positive Metastatic Breast Cancer: A Randomized Clinical Trial. *JAMA* **2017**, *317*, 37–47. [CrossRef] [PubMed]
19. Pivot, X.; Bondarenko, I.; Nowecki, Z.; Dvorkin, M.; Trishkina, E.; Ahn, J.-H.; Im, S.-A.; Sarosiek, T.; Chatterjee, S.; Wojtukiewicz, M.; et al. A Phase III Study Comparing SB3 (a Proposed Trastuzumab Biosimilar) and Trastuzumab Reference Product in HER2-Positive Early Breast Cancer Treated with Neoadjuvant-Adjuvant Treatment: Final Safety, Immunogenicity and Survival Results. *Eur. J. Cancer* **2018**, *93*, 19–27. [CrossRef] [PubMed]
20. Stebbing, J.; Baranau, Y.; Baryash, V.; Manikhas, A.; Moiseyenko, V.; Dzagnidze, G.; Zhavrid, E.; Boliukh, D.; Stroyakovskii, D.; Pikiel, J.; et al. CT-P6 Compared with Reference Trastuzumab for HER2-Positive Breast Cancer: A Randomised, Double-Blind, Active-Controlled, Phase 3 Equivalence Trial. *Lancet Oncol.* **2017**, *18*, 917–928. [CrossRef] [PubMed]
21. Stebbing, J.; Mainwaring, P.N.; Curigliano, G.; Pegram, M.; Latymer, M.; Bair, A.H.; Rugo, H.S. Understanding the Role of Comparative Clinical Studies in the Development of Oncology Biosimilars. *J. Clin. Oncol.* **2020**, *38*, 1070–1080. [CrossRef] [PubMed]
22. Moore, T.J.; Mouslim, M.C.; Blunt, J.L.; Alexander, G.C.; Shermock, K.M. Assessment of Availability, Clinical Testing, and US Food and Drug Administration Review of Biosimilar Biologic Products. *JAMA Intern. Med.* **2021**, *181*, 52–60. [CrossRef] [PubMed]
23. Chow, S.-C.; Liu, J. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2008; ISBN 978-1-58488-668-6.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.