Investigating the Non-Inferiority of a Single Dosage of Curosurf for the Treatment of Neonatal Respiratory Distress Syndrome

Adam Jauregui and Brandon Lunberg

December 18, 2016

1 Introduction

Respiratory Distress Syndrome (RDS) is a breathing disorder that affects premature newborns due to underdeveloped lungs. This condition affects the lungs by keeping them from fully opening due to a lack of the substance that is meant to coat the alveoli. This causes a lack of oxygen in the blood. If an infant goes without treatment of RDS, it can lead to extreme brain damage and even death. Every year in the United States, of the 12% of babies born prematurely, 10% of them will develop RDS (Schraufnagel, 2010). With proper treatment, a significant number of these cases can be cured. A leading rescue treatment for neonatal RDS is Curosurf.

Curosurf is a surfactant that acts as a substitute for the lack of the substance that is needed to coat the alveoli. Curosurf has been proven to reduce mortality in newborns and reduces the harmful air or gas in the pleural cavity. Much research has been done on determining what the optimum dosing regimen of Curosurf should be. In one of the first studies done on Curosurf, a dosage of 200mg/kg was shown to cure 55% of babies as opposed to 26% for a control group that only received supplemental oxygen (Surfactant Replacement Therapy for Severe Neonatal Respiratory Distress Syndrome, 1988). Speer, et. al (1992) compared proportions of a single-dose (200mg/kg) treatment of Curosurf versus a multiple-dose (200mg/kg, then up to two additional doses of 100 mg/kg each) treatment of Curosurf. Halliday, et. al (1993) compared

two multiple dose treatments between a low-dosage group (receiving 100 mg/kg at entry, then up to two further doses of 100mg/kg each) and a high-dosage group (receiving 200mg/kg at entry, then up to four further doses of 100mg/kg each).

In our research, we investigate the non-inferiority (NI) of a single dose regimen (200 mg/kg at entry) of Curosurf to a historical cohort, most specifically from Halliday, et al.'s (1993) low and high dosage group. As per a non-inferiority trial we needed to assign a non-inferiority margin. We decide to test different margins using a synthesis method. In addition, we analyze our primary outcomes under an ordered three-level categorical endpoint setting, as introduced by Brittain and Hu (2009). This latter method is used if we want to assign a value to our intermediate outcome, which we felt was approporiate in our research. Since this value is subjective, we test different values for this as well. We measure the success of our non-inferiority trial by computing the lower bound of a 95% confidence interval.

Our paper is organized as follows. In Section 2, we describe the dataset and the baseline characteristics. Section 3 details the methodology behind non-inferiority trials, how we chose the margin, and assigning a value to our intermediate outcome. In Section 4, we present our results, drawing inference from whether the lower bound of the confidence interval is less than the non-inferiority margin. Section 5 summarizes our findings and possible directions for future work.

2 Dataset

There were three primary outcomes in both the local and historical datasets. The outcomes were measured 28 days after birth for each subject. The outcomes were as follows:

1) Alive without oxygen dependency

2) Alive with oxygen dependency

3) Death

There were a number of babies in our dataset that were not included in our analysis. Of the 317 babies within our local dataset, 34 were listed as "Unknown," meaning none of the primary outcomes were recorded for them. Additionally, there were two babies with data entry errors that were also discarded from our analysis. One of them did not have a data entry record in any

of the primary outcomes nor listed as "Unknown," while the other baby had a recorded entry for both outcomes (1) and (2). This meant that in all 281 babies were included in our analysis.

Listed in Table 1 is a side-by-side comparison of the baseline characteristics between our local dataset and the two groups from the historical datseset. In comparison to the two historical groups, we can see that our local dataset had a higer percentage of females. Additionally, the median in our local dataset was smaller than that of the historical dataset for the variables gestation weeks, gestation weight, and Apgar score. Lower medians for these variables could imply that our dataset is at a disadvantage to the historical groups from the onset. However, with the actual data from the historical groups being unavailable, we have no way of testing whether these differences are significant.

Table 1: Comparison of study groups at entry

Baseline Characteristics	*Local (n=281)*	*Historical Low (n=1069)*	*Historical High (n=1099)*
Male	49.82%	57.90%	57.20%
Average Gestation Weeks	28.05 (2.47)	29.4 (3.1)	29.3 (3.2)
Average Gestation Weight	1093.2 (352.41)	1390 (604)	1358 (606)
Median Apgar Score (1 min.)	4	5	5
Median Apgar Score (5 min.)	7	8	8

3 Methodology

In a standard hypothesis test setting, if we have two treatment groups with one given Drug C (an active control drug) and the other given Drug E (an experimental drug), and we wish to test whether the difference in success rate between Drug C and Drug E is unequal then our null and alternative hypothesis is

$$H_0 : \hat{p}_e = \hat{p}_c$$

$$H_A : \hat{p}_e \neq \hat{p}_c$$

where $\hat{p}_e$ is the proportion of success for Drug E and $\hat{p}_c$ is the proportion of success for Drug C.

However, if the treatment groups above were analyzed in a non-inferiority trial setting, then we would include a pre-specified non-inferiority margin Δ in our null and alternative hypothesis.

This NI margin Δ represents the level of loss in effectiveness we're willing to accept between Drug C and Drug E. There are a number of factors for why we would want to do this, such as to control costs and lessen more harmful side effects. Thus, in the NI setting, our null and alternative hypothesis become

$$H_0 : \hat{p}_e \leq \hat{p}_c - \Delta$$
$$H_A : \hat{p}_e > \hat{p}_c - \Delta.$$

From the setup above, we see that we have significance if the proportion of success for Drug E is greater than the proportion of success for Drug C minus the prespecified margin Δ. Figure 1 contains a forest plot of the different outcomes we can obtain from a non-inferiority trial.

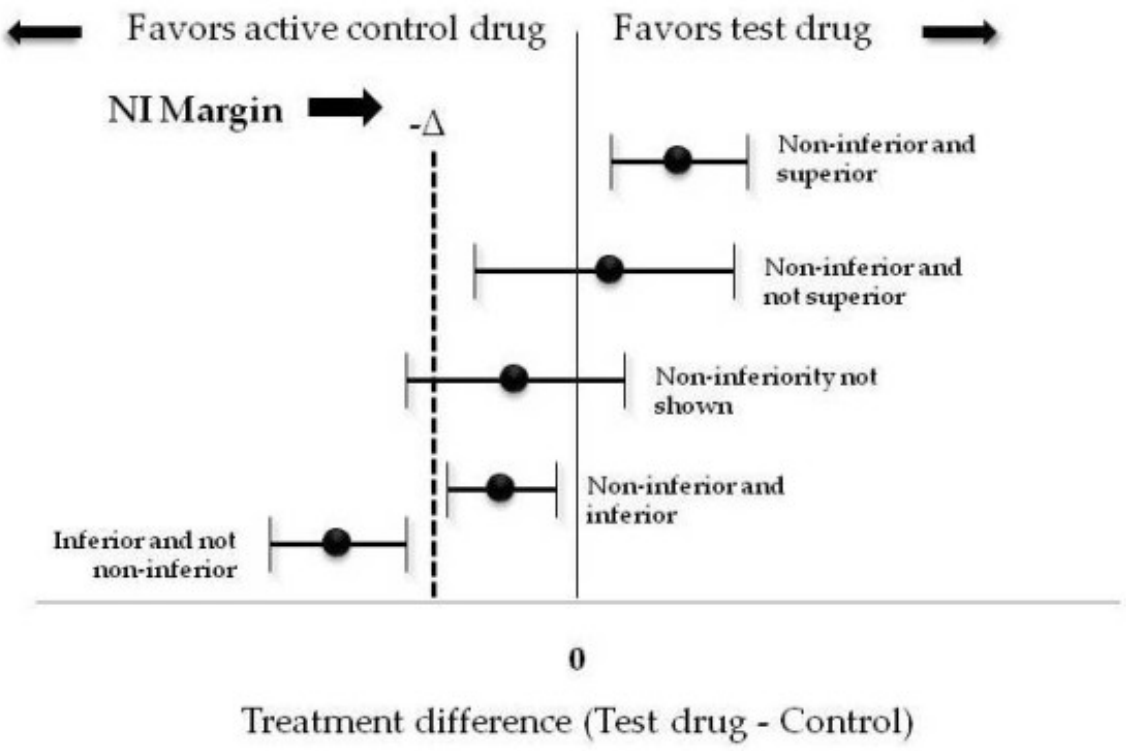


Figure 1: Forest plot outcomes of a non-inferiority trial.

As one can see, this choice of Δ impacts how we make our inferences and requires much justification.

3.1 Choosing the non-inferiority margin

There are many methods to choose the NI margin, such as the Delphic method, the 95-95 method, and the synthesis method (Schumi & Witts, 2011). In the Delphic method, clinical experts who are knowledgeable of the different treatments and disease are consulted and asked for their opinion on how much effectiveness the alternative drug should be allowed to lose. In the 95-95 method, one computes the entire effect of the active control drug relative to a meta-analysis on all the historical placebo trials and a 95% confidence interval is obtained around the estimated difference between the active control and placebo. Denote the lower limit of the confidence interval M_1. Our margin, denoted M_2, is then chosen to preserve a certain percentage of the effect between M_1 and the active control drug, say 50% or 75%. In the synthesis method by contrast, we simplify specify a certain percentage of the active control treatment that we're willing to preserve but no less. This is the method we use in our research.

In choosing our margin, we test three different Δs at 99%, 95%, and 90% of the success of the active control treatment. As an example, Figure 2 lists the percentages of each of our primary outcomes for each of the three groups.

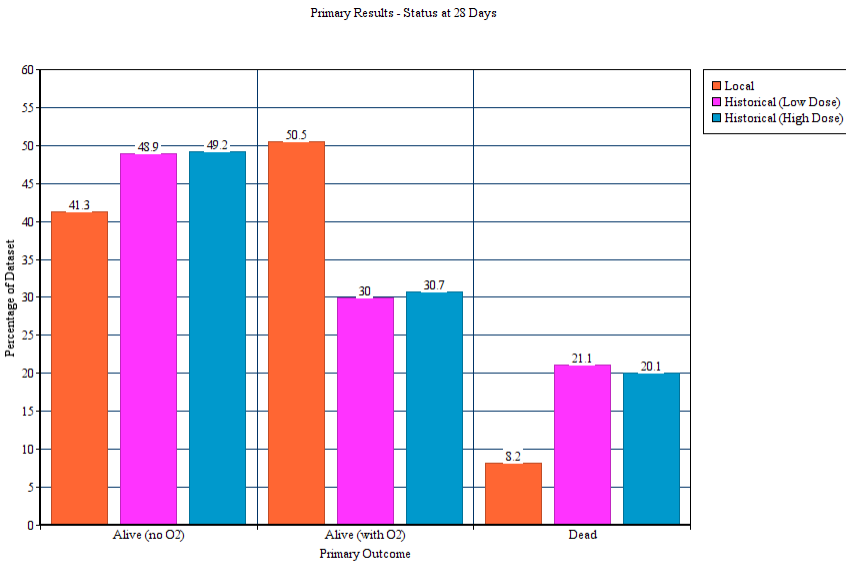


Figure 2: Bar chart of primary outcomes across the single, low-multiple, and high-multiple dosages of Curosurf.

Thus, if success was defined as the proportion of outcome (1), then a Δ that preserves 90%

of the effect of the historical low-multiple dosage group would give a value of .0489 for Δ, so given that the proportion of success in our single dosage was .414 and that the proportion of success in the low-multiple dosage was .489, then our null and alternative hypothesis would be

$$H_0 : .414 \leq .489 - .0489$$

$$H_A : .414 > .489 - .0489.$$

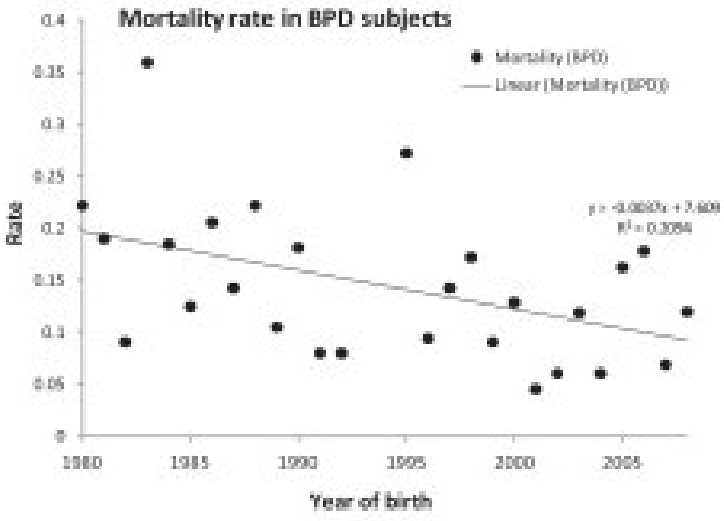
We felt this was a generous margin to have, as we found that many of the past studies done on Curosurf with single dosages had success rates in the 50's (Surfactant Replacement Therapy for Severe Neonatal Respiratory Distress Syndrome, 1988), so we did not want our success rates too far off from that study.

3.2 Interpreting the intermediate outcome

In past research on Curosurf, such as with the first Curosurf study done in 1988, Speer, et al. (1992), and Halliday et al. (1993), they share three primary outcomes exactly as ours: (1) alive without oxygen dependency, (2) alive with oxygen dependency, and (3) death. However, in their statistical analysis, their categorical endpoints are binary, where "success" is defined as the proportion of outcome (1) and "failure" is defined as proportion of the outcomes (2) and (3) grouped together. We argue against the grouping of outcomes (2) and (3). One reason for this is because outcome (2), "oxygen dependency" or bronchopulmonary dysplasia (BPD), is only associated with a mortality rate between 5% and 25% for the past thirty years, with the trend pointing downward as shown in Figure 3 below taken from Zysman-Colman et. al's (2013) research.

These researchers are not denying that BPD is a serious disease and should only be measured by its mortality rate, but it seems unfair to simply group BPD and death together, especially in the context of a non-inferiority trial, where we hope to show non-inferiority for a drug that might have fewer side effects (such as the rate of death). To remedy this, we use a methodology introduced by Brittain and Hu (2009).

For our analysis, we assign a value of 1 for the success outcome (alive without oxygen dependency), 0 for the failure outcome (death), and a subjective value denoted by ρ for the intermediate outcome (alive with oxygen dependency). As outlined by Brittain and Hu (2009), this value of ρ is purely subjective and its value is up to the researcher. If the researcher feels that



the intermediate outcome is more of a success than failure, then its value might be .75. Similarly if the intermediate outcome is thought to be more of a failure than success. However, they argue assigning ρ a value of .5 is the most natural interpretation of the intermediate outcome. In our research, we test different values of ρ: 0, .5, .75, and 1.

4 Results

In Table 2 presented below are the secondary outcomes resulting from the dosages of Curosurf in the local and historical dataset. Our local dataset fared worse within the categories of duration of more than 40% O2 (days), duration of intubation, and the onset of sepsis. However, the single dosage of Curosurf had lower incidences of pneumonia, necrotizing enterocolitis, and pulmonary hemorrhage. These results are encouraging as they provide a sound basis for choosing to do a non-inferiority trial on our dosage of Curosurf.

When testing the non-inferiority of the drug in comparison to the historical cohort, we analyze the upper and lower bounds to get an idea of the drugs effectiveness under each value of ρ. The focus is mainly on the lower bound (LB) of the confidence interval.

$$LB = \hat{p}_e - \hat{p}_c - z_{1-\alpha}\sqrt{\frac{\hat{p}_e(1-\hat{p}_e)}{n_e} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_c}}$$

Table 2: Comparison of study groups post-treatment

Secondary Results	*Local (n = 281)*	*Low Dose (n = 1069)*	*High Dose (n = 1099)*
Duration of >40% O2 (days) (avg)	4.9	3	3
Duration of intubation (avg)	6.9	6	5
Pneumonia	5.33%	10.60%	12.20%
Sepsis	28.46%	20.20%	21.40%
Necrotizing Enterocolitis	2.84%	5.30%	6.70%
Pulmonary Hemorrhage	2.84%	5.60%	6.80%

The results of the non-inferiority test are determined by the following cases:

$$\begin{aligned} LB < -\Delta &: \text{Not Non-Inferior} \\ -\Delta < LB < 0 &: \text{Non-Inferior} \\ LB > 0 &: \text{Superior} \end{aligned}$$

We can obtain statistical significance in the test if the lower bound is greater than $-\Delta$. However, since we have assigned our intermediate outcome a value represented by ρ, we must use a modified version of the lower bound:

$$LB = ((\hat{p}_{se} + \rho\hat{p}_{ie}) - (\hat{p}_{sc} + \rho\hat{p}_{ic})) - z_{1-\alpha} \sqrt{\frac{\hat{p}_{se}(1-\hat{p}_{se}) + \rho^2(\hat{p}_{ie}(1-\hat{p}_{ie})) - 2\rho\hat{p}_{se}\hat{p}_{ie}}{n_e} + \frac{\hat{p}_{sc}(1-\hat{p}_{sc}) + \rho^2(\hat{p}_{ic}(1-\hat{p}_{ic})) - 2\rho\hat{p}_{sc}\hat{p}_{ic}}{n_c}}$$

where $\hat{p}_{se}, \hat{p}_{ie}, \hat{p}_{sc}$, and $\hat{p}_{ic}$ represent respectively the proportion of success for the experimental drug, the proportion of the intermediate outcome in the experimental drug, the proportion of success for the active control drug, and the proportion of the intermediate outcome in the active control drug. Notice that when ρ=0, then this is exactly the same as the lower bound equation as presented earlier where we group outcomes (2) and (3). And when $\rho = 1$, this is equivalent to grouping outcomes (1) and (2) together. The confidence interval was determined using a level of significance of α=.05.

From Figure 4, it is shown that as the value of ρ increases, the results favor the experimental dose over both historical datasets. When the value of ρ is at least .5, all but one test case for each of the two doses showed non-inferiority. In the last test, where the value of $\rho = 0$, we can

not prove non-inferiority, as the lower bound is much less than all three non-inferiority margins. These results confirm that the effectiveness of the drug relies on the interpretation of the harm of subjects with BPD.

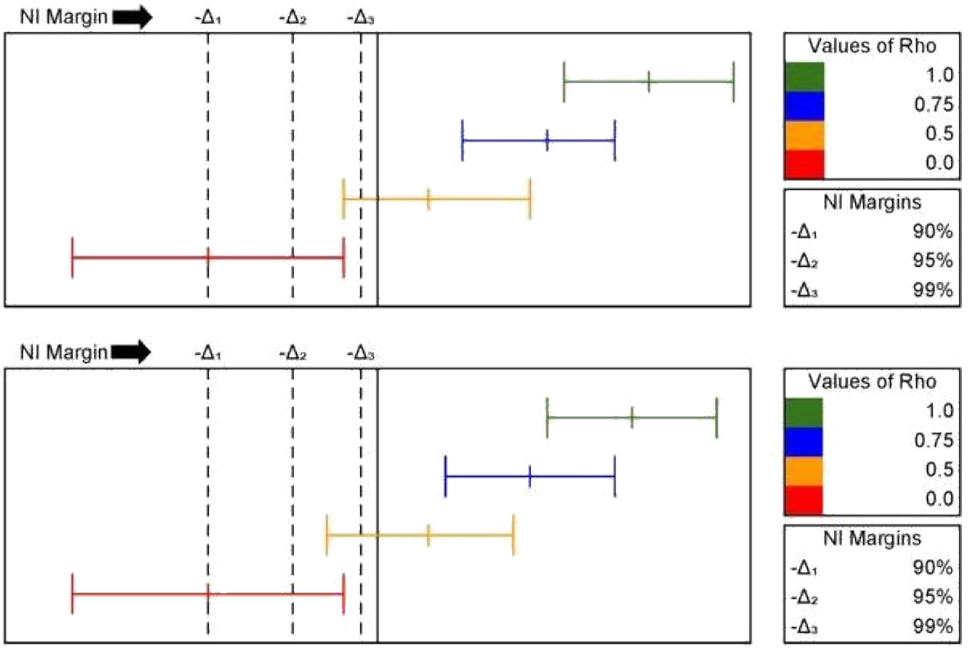


Figure 3: Confidence intervals for the experimental treatment against the historical low dose treatment (1) and the historical high dose treatment (2).

5 Conclusion and Discussion

The goal of this research was to compare the study done out of Community Regional Medical Center in Fresno, CA against a historical dataset to show that the new dose could be a substitute for the current Curosurf treatment dose. The historical dataset focused on the success of the drug being reliant on the subject being alive without oxygen dependency 28 days after birth. When comparing the experimental data against the historical data in this way, the experimental data did not meet the same success rate as the historical data. However, if the interpretation of a subject with oxygen dependency is changed a bit, the results show something different. Due to the relatively low concerns with subjects that are oxygen dependant, it can be argued that combining oxygen dependency with death may not be the correct approach in looking at the

success of this drug. The experimental data showed a significantly lower mortality rate. If it is true that most subjects that are oxygen dependent end up with a clean bill of health long term, such as suggested by the American Lung Association, it seems as if those subjects should not be grouped with the dead subjects. It may even be plausible that they lean more towards a success than a failure.

For future work on the subject, it would be interesting to see how an expert in the field of breathing disorders would come up with the value for the intermediate primary result, as well as deciding the non-inferiority margin for the test. Tweaking these two values have a huge effect on the results of the test using our model, which could ultimately determine the effectiveness of the drug.

6 References

1. Brittain, Erica, and Zonghui Hu. Non-Inferiority Trial Design and Analysis with an Ordered Three-Level Categorical Endpoint. Journal of biopharmaceutical statistics 19.4 (2009): 685699. PMC. Web. 12 May 2016.

2. Halliday H, Tarnow-Mordi W, Corcoran J, Patterson C. Multicentre randomised trial comparing high and low dose surfactant regimens for the treatment of respiratory distress syndrome (the Curosurf 4 trial). Arch. Dis. Child. 1993;69(3):276280.

3. Schraufnagel, Dean E. "19." Breathing in America: Diseases, Progress, and Hope. New York: American Thoracic Society, 2010. N. pag. Print

4. Schumi, Jennifer, and Janet T. Wittes. "Trials." Through the Looking Glass: Understanding Non-inferiority. N.p., 3 May 2011. Web. Mar. 2016.

5. Speer CP, Robertson B, Curstedt T, Halliday HL, Compagnone D, et al. (1992) Randomized European multicenter trial of surfactant replacement therapy for severe neonatal respiratory distress syndrome: single versus multiple doses of Curosurf. Pediatrics 89: 1320

6. Surfactant Replacement Therapy for Severe Neonatal Respiratory Distress Syndrome: An International Randomized Clinical Trial Pediatrics Nov 1988, 82 (5) 683-691

7. Zysman-Colman, Zofia et al. Bronchopulmonary Dysplasia Trends over Three Decades. Paediatrics & Child Health 18.2 (2013): 8690. Print.