A Bayes Factor Approach to Noninferiority Trials

Introduction

Before new drug treatments are approved for public use, they must be tested for safety and efficacy through a series of clinical trials. Observational studies come with bias, so it's difficult to directly attribute any difference in treatments to a particular treatment. Therefore, investigators need to employ experimental design principles in clinical trials to achieve more rigorous evidence of effectiveness. A common controlled clinical trial design is the superiority trial, where the intent is to show that the new treatment is more effective than the current standard or placebo. Once data is collected from the treatment groups of interest, practitioners will usually conduct a t-test to assess if there a statistically significant difference in the group means (Cook & Demets, 2008).

However, it's not ethical to include a placebo as a control if it means that patients will suffer unnecessary physical or mental suffering. Additionally, a superiority trial might be a waste of resources when there are only ancillary benefits, such as decreased drug production cost or increased patient adherence. In these situations, practitioners look to noninferiority trials designs, which can be complicated to design. Besides ensuring that a sufficiently powered study through sample size, there are other critical decisions such as selecting the "best" active control for comparison (D' Agostino Sr., Massaro, & Sullivan, 2003). In the pursuit of showing that a new treatment is not inferior to the active control, investigators are making two different comparisons (Dmitrienko, Tamhane, & Bretz, 2009):

- Direct comparison between the treatment and active control
- Indirect comparison of treatment against a placebo

Conclusion of noninferiority is based on a prespecified zone of noninferiority. If the effect size falls within this margin, then the groups of treatment could be considered similar. When superiority trials are conducted poorly or are underpowered, then the worst outcome is that results may be biased

towards a treatment not being significantly better than a placebo or active control. However, when it comes to NI trials, several scenarios could result in a false conclusion of noninferiority (Cook & Demets, 2008). Some influences include (Dmitrienko, Tamhane, & Bretz, 2009):

- treatment differences are reduced due to patient nonadherence
- the least effective control group was chosen for comparison
- standard care has changed over time, affecting the estimate of the active control

Biocreep occurs when a new, inferior treatment is falsely accepted as "efficacious" after a series of NI trials. The long-term consequence is that the pool of treatments to gradually become less effective and therapies that are no better than a placebo are approved (Everson-Stewart & Emerson, 2010).

Methods

The NI Margin

Before data collection or analysis happens, the noninferiority margin must be predefined. Usually based on practical clinical significance, the NI margin is the fraction of the active control's true effect that the new therapy is allowed to be worse than. Additionally, the NI margin cannot be greater than the smallest effect size that the active control would be expected to have if it had been tested in superiority trials (D' Agostino Sr., Massaro, & Sullivan, 2003). The most common method for determining the NI margin is called the fixed margin method, or the 95%-95% method. Assuming that the effect of the active control is constant across studies, the calculation is performed like so (CDER & CBER , 2016):

1. Estimate the active control effect with past placebo-controlled studies.
2. Define M_1, which is usually the lower bound of the 95% CI (confidence interval) for the estimated active control effect.
3. Determine M_2, which is usually set to be at least 50% of M_1.

An alternative method is the synthesis method, which combines the estimate of treatment effect relative to the control from the current NI trial with the estimate of the control effect from the meta-analysis of the historical trials. The con with this approach is that it's not possible to use choose the NI margin in advance of the trial. This has ramifications for sample size calculations, which are crucial in ensuring that studies are properly powered (CDER & CBER , 2016).

The Frequentist Analysis

Once the study has been conducted, investigators can conduct a one-sided t-test for the differences in group means or proportions with the following hypotheses, where the null hypothesis is that that the control is inferior to the treatment by the NI margin or more:

$$H_0: \theta_T - \theta_C < -M_2$$

$$H_A: \theta_T - \theta_C \geq -M_2$$

If the effect size δ and the 95% CI calculated from the observed data sits completely above the prespecified NI margin, then that trial has shown non-inferiority between the control and treatment. However, there are many possible results, as shown in Figure 1, where investigators could have an inconclusive conclusion (Schumi & Wittes, 2011).

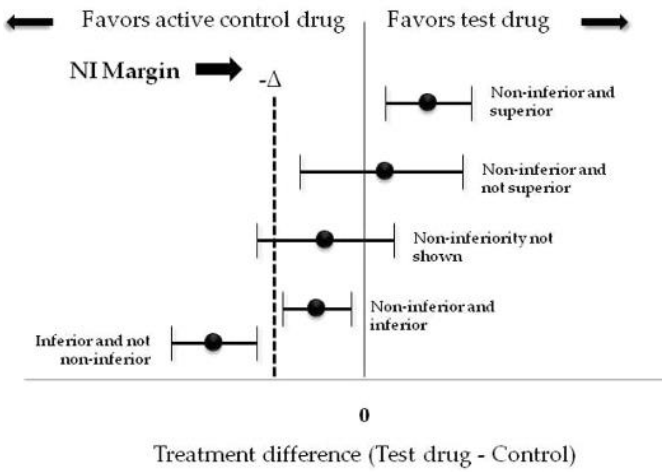

Figure 1. Chart displaying different CI results if the study was assessing positive outcomes.

Bayesian Approach

When it comes to null hypothesis significance testing, a non-significant p-value could have been the result of H_0 being true or a combination of the H_0 being false and an underpowered trial (van Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2019). Unlike the frequentist method, the Bayes factor

(BF) hypothesis testing based is capable of summarizing evidence in favor of the null hypothesis relative to the alternative hypothesis. This is helpful when the goal is determining the presence or absence of a treatment effect and can help prevent wasted resources (Stefan, Gronau, Schonbrodt, & Wagenmakers, 2019). Though it could involve potentially extensive computation time, some benefits of BFs include a reduction in sample sizes while maintaining sufficient power and the ability to incorporate historical data directly in the analysis (Li, Chen, & Wang, 2018).

The BF is a relative metric that quantifies the modification of the prior odds of H_1 over H_0 because of the observed data *y* (Stefan, Katsimpokis, Gronau, & Wagenmakers, 2022):

$$BF_{10}(d) = \frac{p(y|H_1)}{p(y|H_0)} = \frac{\frac{p(H_1|y) * p(y)}{p(H_1)}}{\frac{p(H_0|y) * p(y)}{p(H_0)}} = \frac{p(H_0)}{p(H_1)} * \frac{p(H_1|y)}{p(H_0|y)}$$

Another way to interpret the BF is that it's the ratio of marginal likelihoods averaged across all possible posterior values of θ under the two sets of hypotheses. Note that B_{10} can be interpreted as quantitative evidence in favor of H_1 over H_0. For example, $BF_{10} = 8$ is considered as moderate evidence since the data is 8 times more likely under H_1 compared to H_0. Generally, a BF_{10} between 1 and 3 is considered not significant evidence, between 3 and 20 is considered some positive evidence, and greater than 20 is considered strong evidence.

Tests that involve BFs do depend on having a proper prior since improper prior distributions would result in too many competing solutions in the case of two-sided tests (Robert, 2007). If investigators have no prior knowledge when performing an informed t-test, then it is suggested they default to a Cauchy distribution with location parameter = 0, scale parameter = 1, and degrees of freedom = 1. This prior fulfills the requirements of information consistency and predictive matching that are desired in the BF hypothesis framework. If the data is completely uninformative or if the sample size is insufficient, then the predictive matching condition expects a result of $BF_{10} = 1$, which indicates that there is no evidence to persuade us in either direction of the hypotheses. If the data is overwhelmingly

informative against the null hypothesis, then the BF_{10} is information consistent since the value should tend towards to infinity (Ly, et al., 2020).

Case Study Reanalysis

To explore the limits of the BF hypothesis framework, I am replicating van Ravenwaaij et al's work. The data utilized comes from a cluster-randomized crossover trial whose purpose was to test the noninferiority of the beta-lactam strategy to the beta-lactam-macrolide and fluoroquinolone strategies in treating clinically suspected community-acquired pneumonia (CAP). The original study that produced the data, led by Postma et al, was motivated by how guidelines in the Netherlands increased the use of macrolides and fluoroquinolones as an antibiotic treatment, even though there was limited evidence; the effects of beta-lactam-macrolide were not validated through carefully controlled superiority trials and the recommendation for fluoroquinolone were based on retrospective observational studies and the convenience of a once daily dosing treatment.

The primary measure of interest is 90-day mortality and differences between the treatment groups were analyzed from an intention-to-treat perspective. An NI margin of 3 percentage points was used when assessing the 2-sided 90% CIs. Based on the observed data below, and with respect to the beta-lactam strategy, the researchers determined the following (Postma, et al., 2015):

Treatment	Mortality Count	Sample Size	Mortality Rates (%)
Beta-lactam (B)	59	656	9.0
Beta-lactam-macrolide (BM)	82	739	11.1
Fluoroquinolone (F)	78	888	8.8

- Mortality was higher with beta-lactam-macrolide by 1.9 % with a 90% CI of (-0.6, 4.4)
- Mortality was lower with fluroquinolone by 0.6 % with a 90% CI of (-2.8, 1.9)

Since the calculated CIs do not include the prespecified NI of 3%, the researchers determined that the beta-lactam strategy was noninferior to the other treatments. One aspect of the study that was

unusual was that the investigators did not accept the analysis with 95% CIs. As shown in Figure 2, there would have been inconclusive results about noninferiority when comparing beta-lactam to

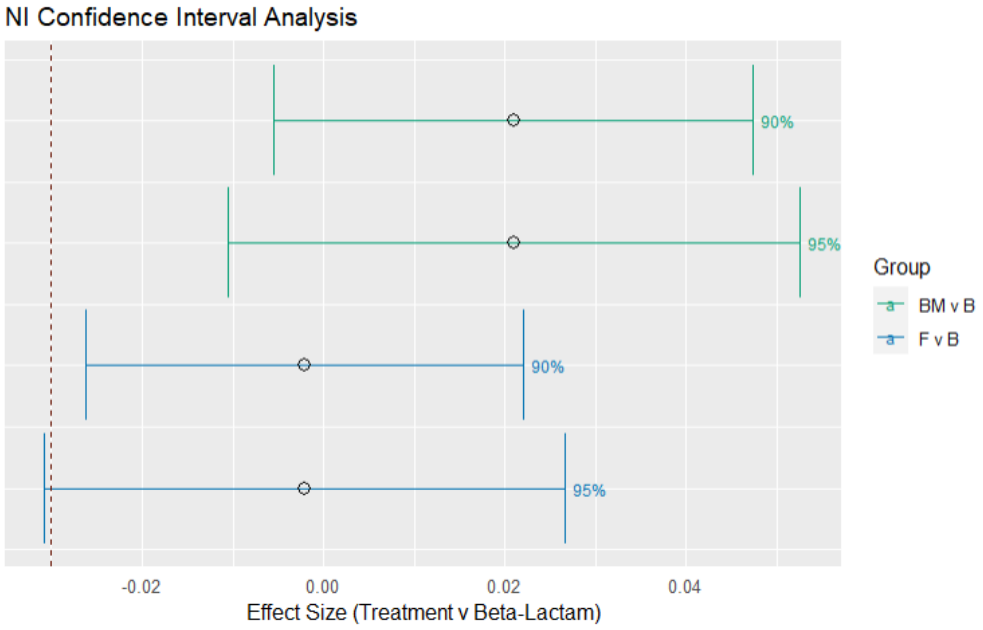

Figure 2. The estimated 90% and 95% CIs for each of the treatment group comparisons. Group 1 compares beta-lactam-macrolide to beta-lactam, while Group 2 compares fluoroquinolone to beta-lactam.

fluoroquinolones. Postma et al justified the use of the 90% CIs since they did not notice a clear trend favoring fluoroquinolones and since beta-lactam monotherapy wasn't associated with longer hospital stays or higher rates of complications in other analyses that they performed (Postma, et al., 2015).

Informed T-Test Analysis

Since the estimated mortality proportions of beta-lactam (p_B) and fluoroquinolone (p_F) are very close to each other, it's more difficult to assess if the effect size ($\delta = p_F - p_B$) is statistically significant. Additionally, when it comes to evaluating the difference in group proportions, the absolute difference doesn't directly translate to a meaningful difference. For example, the difference between 0.01 to 0.02 % is more interesting because the effect size has doubled compared to the difference between 0.56 to 0.57 %. To effectively communicate practical significance, the difference in proportions between the treatment groups is converted into Cohen's *h*, which is a standardized effect size. The closer the Cohen's

h value is to 1, the larger the magnitude of practical significance. Researchers can then compare the effect sizes across studies; Cohen's *h* is considered small if 0.2, medium if 0.5, and large effect if 0.8 (Dey & Mulekar, 2018). In the noninferiority setting, we make use of the constraint $p_F - p_B = -M_2$ in calculating Cohen's *h*. To simplify the calculation, I will utilize the pooled proportion values but ensure that the total shift in distribution is equal to the value of the NI margin. Instead of simply subtracting the value from the pooled proportion one-sidededly like van Ravenzwaaij et al, I will ensure symmetry by subtracting and adding $\frac{-M_2}{2}$ from the pooled proportion. Doing so results in the following Cohen's *h* formula:

$$h = 2\left[arcsin(\sqrt{p_F}) - arcsin(\sqrt{p_B})\right] = 2\left[arcsin\left(\sqrt{p_{pooled} - \frac{M_2}{2}} - arcsin(\sqrt{p_{pooled} + \frac{M_2}{2}})\right)\right]$$

Next, the hypotheses being compared must be carefully defined for the NI case study. First, we assume that the null hypothesis has a positive point mass so that we can quantify how much it's supported or undermined by the data. In this case study, the primary measure is mortality, a negative outcome. If δ < 0, then beta-lactam results in higher mortality and should be considered inferior. When translating the NI trial's question of inferiority, the following sets of point-null hypotheses result and will be used to obtain BF_{-0} and BF_{0+}. Respectively, the sets represent the scenario where beta-lactam is inferior or noninferior:

$$H_0: p_F - p_B = -M_2 \qquad H_0: p_F - p_B = -M_2$$

$$H_A: p_F - p_B < -M_2 \qquad H_A: p_F - p_B > -M_2$$

In computing the BF, we're computing the posterior probabilities under two composite hypotheses. Using the principal of transitivity, the significance of how likely the data *y* is to be observed under the two hypotheses can be calculated through:

$$BF_{+-} = \frac{BF_{0-}}{BF_{0+}} = \frac{\frac{p\,(y|\,\theta_F - \theta_B = -M_2)}{p(y|\,\theta_F - \theta_B < -M_2)}}{\frac{p\,(y|\,\theta_F - \theta_B = -M_2)}{p(y|\theta_F - \theta_B > -M_2)}} = \frac{p(y|\theta_F - \theta_B > -M_2)}{p(y|\theta_F - \theta_B < -M_2)}$$

To assess the significance of the effect size, the best test is an informed t-test that was extended from Jeffrey's original Bayesian t-test to a two-sample set-up. First, the means of two groups was reparametrized in terms of a grand mean and an effect size. Next, a right Haar prior was imposed on the nuisance parameters common to both the null and alternative hypothesis to help make the BF invariant. By doing so, the marginal likelihood of the null model becomes proportional to the density of a standard t-distribution, evaluated at the observed t-value. Lastly, the prior under H_A was decomposed from $\pi_1(\mu, \sigma, \delta)$ to $\pi_0(\mu, \sigma)\pi(\delta)$ which helps to simplify the BF calculations by factoring out the prior related to H_0. Note that instead we have $\pi(\delta) = \frac{1}{\gamma} T_\kappa\left(\frac{\delta - \mu_\delta}{\gamma}\right)$, which is a flexible t-prior that allows investigators to incorporate expert knowledge about a standardized effect size. With these modifications and some algebra, the BF for an informed two sample t-test is of the form (Gronau, Ly, & Wagenmakers, 2020):

$$BF_{10}(data; \mu_\delta, \gamma, \kappa) = \frac{\int T_\upsilon\left(t \mid \sqrt{n_\delta}\delta\right) \frac{1}{\gamma} T_\kappa\left(\frac{\delta - \mu_\delta}{\gamma}\right) d\delta}{T_\upsilon(t)}$$

where the important parameters are effect size (δ), prior location (μ_δ), prior scale (γ), and prior degrees of freedom (κ). The location parameter determines the center of the t-distribution. The scale parameter determines how the probability is distributed over the t-distribution; for example, a high scale value translates to a wider prior distribution where there is more weight in the tails.

Like the original reanalysis, I will use 1 degree of freedom in my exploration. Since the sample sizes for beta-lactam and fluoroquinolone were sufficiently large to justify approximating the difference of proportions with a normal distribution, a z-statistic could be used instead. However, I will deviate from van Ravenzwaaij et al's usage of the pooled standard deviation since I do not believe that the variance between treatment groups is equal. Therefore, my z-statistic will be of the form:

$$Z = \frac{(p_F - p_B) + M_2}{\sqrt{\frac{p_F(1 - p_F)}{n_F} + \frac{p_B(1 - p_B)}{n_B}}}$$

I used the suggested Cauchy prior, but with the location parameter shifted by the calculated Cohen's *h* so that I could treat the prior as if it was centered at "0". The prior and posterior distributions can be compared in Figure 3 below.

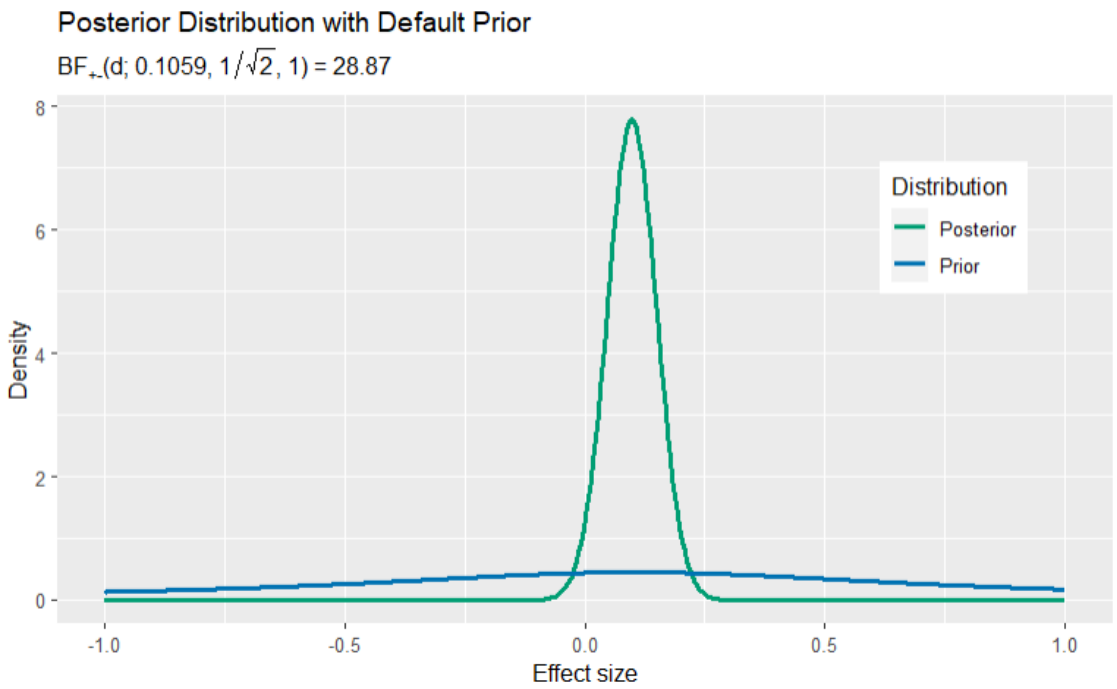

Figure 3. The posterior distribution for the effect size of beta-lactam versus fluoroquinolone is compared to the default Cauchy prior.

Despite using the same code as van Ravenzwaaij et al's, my BF was calculated to be 28.7, while theirs was reported to be 39.07. After making the modifications to calculating the standard error and Cohen's *h* as previously mentioned, my code then calculated a BF of 28.87. Despite the differences, these results corroborate Postma et al's conclusion that beta-lactam is noninferior to fluoroquinolone. A BF value greater than 10 and indicates that there is strong evidence for the control being noninferior, especially since the value indicates that the data observed is about 28 times more likely under the noninferiority hypothesis.

Robustness Check: Choice of Prior Distribution

However, the choice of a prior distribution could impact inferences since the Bayes factor can be considered as a ratio of two prior-weighted averaged likelihoods. Uninformative priors are more easily influenced by data, while strongly informative ones may be more resistant (Gronau, Ly, & Wagenmakers, 2020). Therefore, I will perform a robustness check to assess whether the choice of prior significantly changes the direction of noninferiority conclusions based on the BF values. Based on Figure 4, it looks like the information in the data overwhelms the prior since the prior has little influence on the shape of the posterior distribution. This means that, under the set of hypotheses being considered, the BF result is robust against choice of prior.

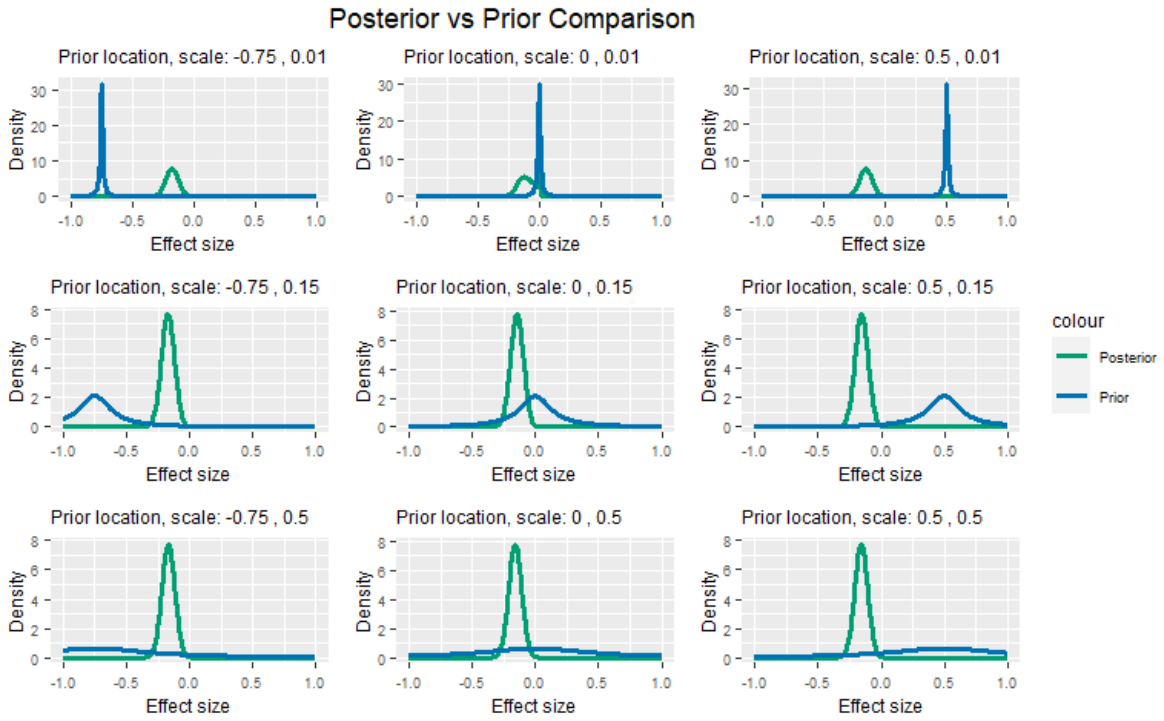

Figure 4. Different prior hyperparameters were specified. As seen, the posterior distribution is not influenced when different hyperparameters are specified for the prior distribution.

To further confirm that the prior makes little impact in this case study, I visualized the BF results in a contour plot, seen in Figure 5. Given the data, the evidence for noninferiority is becomes less compelling below $\log(BF_{10}) = 1$, which corresponds to a $BF_{10} = e = 2.718$. Meanwhile, the evidence for

noninferiority is inconclusive below $\log(BF_{10}) = 0$, which corresponds to a $BF_{10} = 1$ and indicates that there isn't any evidence to prefer the noninferiority or inferiority conclusion. As the prior scale decreases and prior location shifts to the left, the BF increases in favor of noninferiority. For most of the ranges in location and scale values, the direction of the BF test is in favor for concluding that beta-lactam is noninferior to fluoroquinolone.

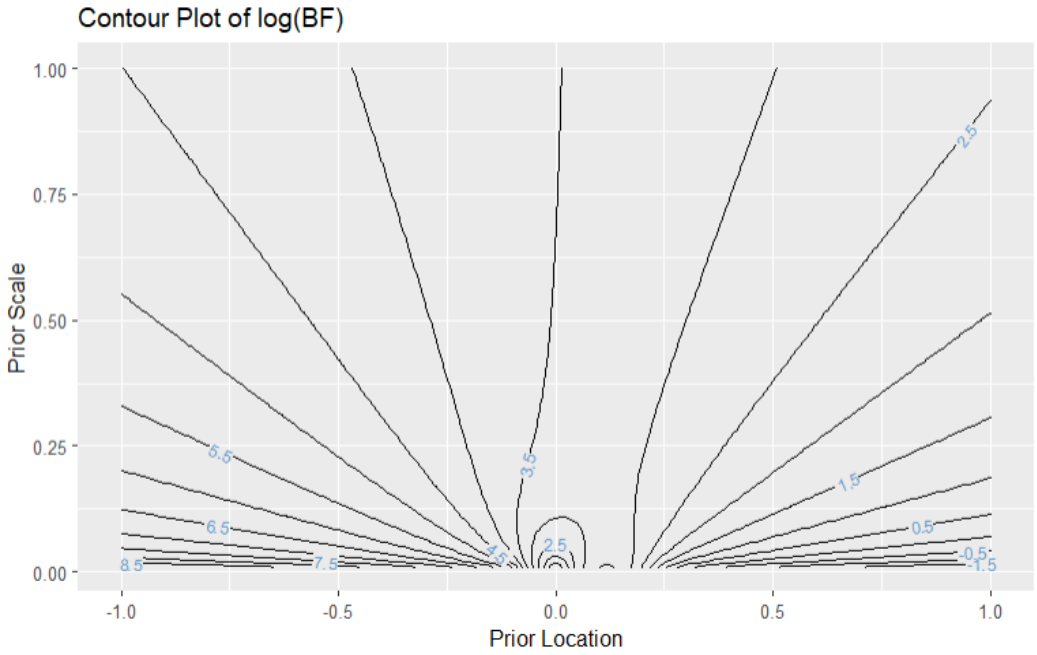

Figure 5. Due to the large values of BF, the density is in terms of log(BF). There is more support for noninferiority based on range of BF values.

Robustness Check: Distribution of True Effect Size

Another robustness check is to assess how the true effect size may impact our decision for noninferiority. To explore this, I will assess the boundaries of BF_{+-} and z-statistic while varying the mortality count between the beta-lactam and fluoroquinolone groups. In the case of a frequentist NI test, we can conclude that the beta-lactam is noninferior when there is significant evidence (z-statistic is greater than 2) that the δ is greater than the NI margin.

As seen in Figure 6, evidence for beta-lactam being noninferior is more likely to result when the mortality count in the beta-lactam group is less than the mortality count in the fluoroquinolone group.

This trend can also be seen when the effect size trends are overlayed on the contour plot in Figure 7.

Size effects that are greater than 0.05 percentage units results in a decision for noninferiority. The larger

δ re is

lit r

n

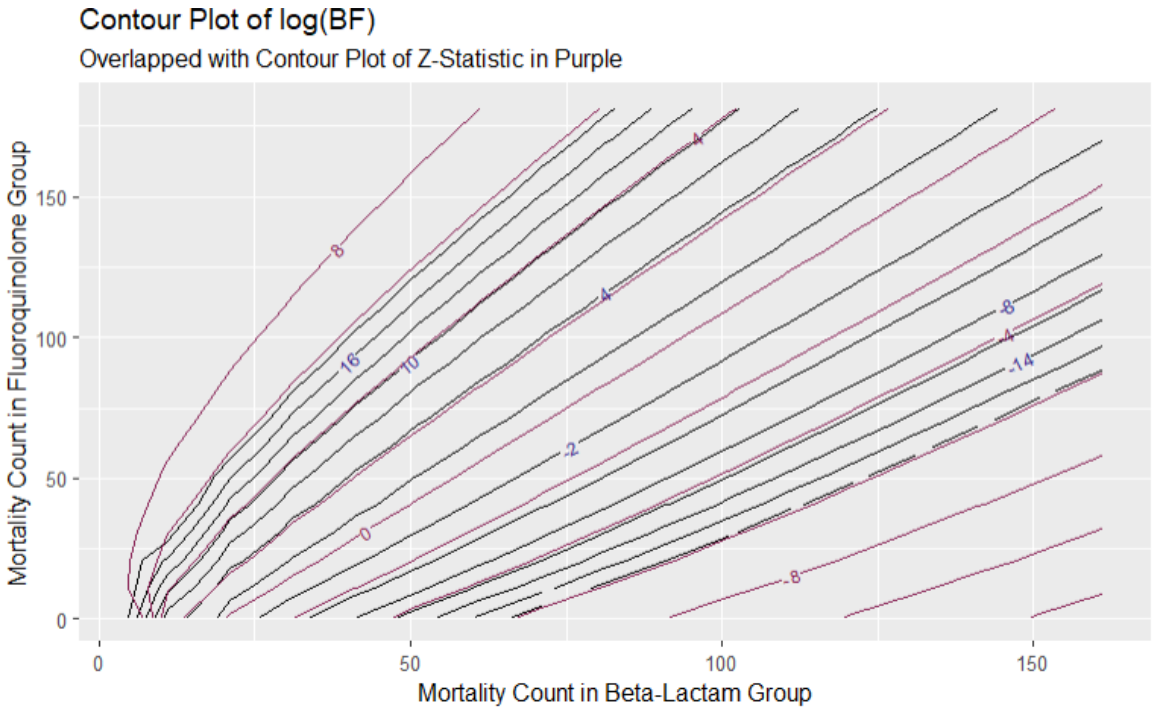

Figure 6. The z-statistic contour plot lines are in purple. The z-statistic = 2 curve aligns closely with log(BF) = 4, which is when there is some evidence for noninferiority.

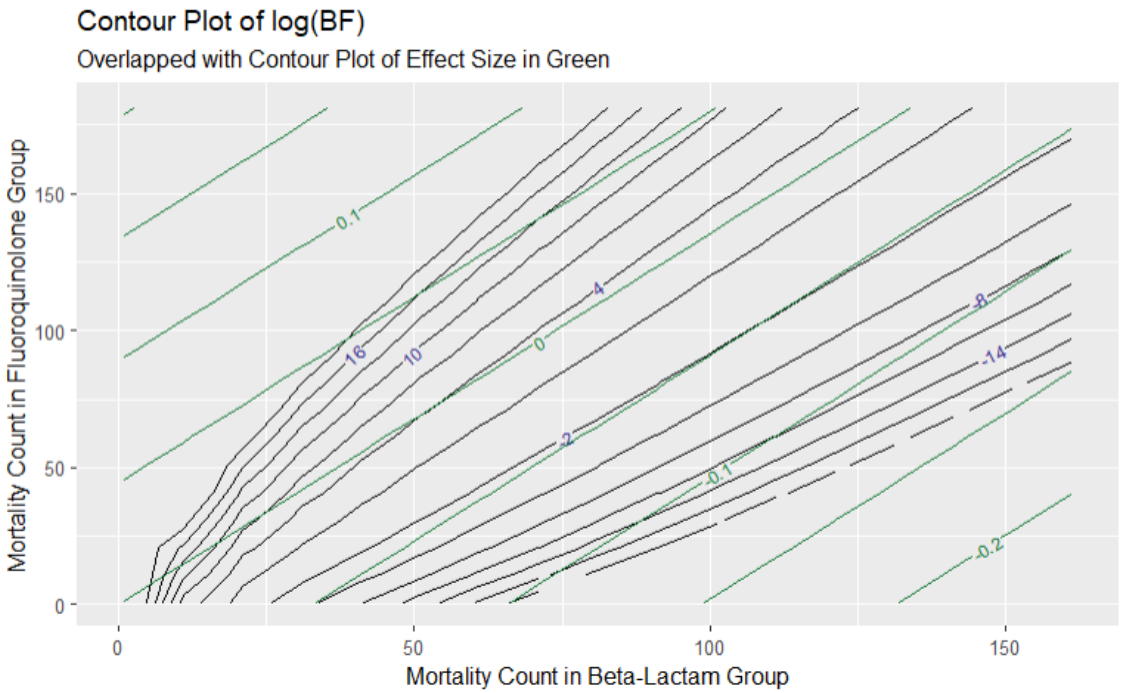

Figure 7. Effect size contour plot lines are in green. Despite the potential for size effect to be small, the BF testing method can detect evidence for noninferiority.

Lastly, as seen in Figure 8 with the decision regions for both methods being shaded, there is no conflict in the direction of the evidence for noninferiority between the two methods. However, it's more likely to conclude noninferiority with the BF method as evidenced by the larger blue shaded region. The frequentist method also does not seem to provide any conclusion in the lower left corner of the plot, where the mortality count in the treatment group is slightly more than a very low mortality count in the control group.

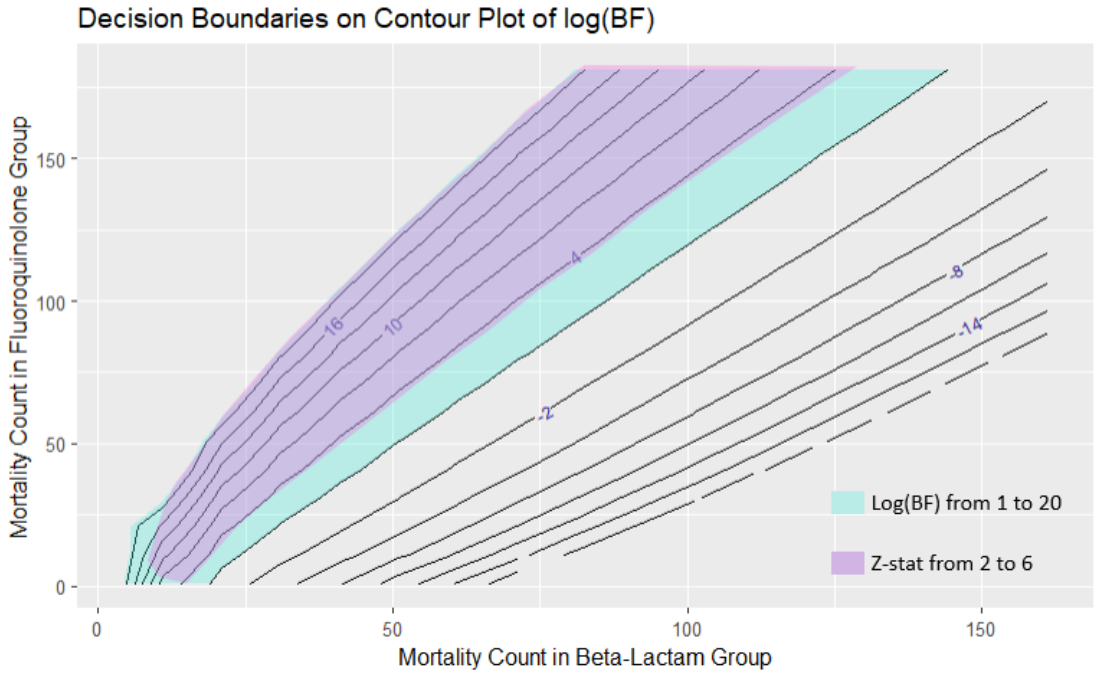

Figure 8. The acceptance range for H_0 from the Frequentist decision to accept that the control is noninferior is colored purple, from $2 < Z < 8$. Meanwhile, the acceptance range for H_0 from the BF perspective is in light blue for $\log(BF) > 1$.

Conclusion

Based on my exploration with a particular case study, there seemed to be little difference in the decision boundaries for non-inferiority between the NHST and BF hypothesis methods. With criticism about how *p*-values are prone to overestimating effects and false interpretations (Kelter, 2020), Bayesian approaches such as the BF hypothesis test can complement research efforts, especially with proof-of-concept studies that have limited resources to spare. Information resulting from such a study would be useful in guiding investigators with where to proceed with their research.

References

CDER & CBER . (2016). *Non-Inferiority Clinical Trials to Establish Effectiveness: Guidance for Industry.* Article FDA-2010-D-0075.

Cook, T. D., & Demets, D. L. (2008). *Introduction to Statistical Methods for Clinical Trials.* Chapman & Hall/Crc.

D' Agostino Sr., R. B., Massaro, J. M., & Sullivan, L. M. (2003). Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine*, DOI: 10.1002/sim.1425.

Dey, R., & Mulekar, M. S. (2018). Effect Size as a Measure of DIfference Between Two Populations. In R. Alhajj, & J. Rokne, *Encyclopedia of Social Network Analysis and Mining* (pp. https://doi.org/10.1007/978-1-4939-7131-2). Springer New York.

Dmitrienko, A., Tamhane, A. C., & Bretz, F. (2009). *Multiple testing problems in pharmaceutical statistics .* CRC Press.

Everson-Stewart, S., & Emerson, S. (2010). Bio-creep in non-inferiority clinical trials. *Stat Med*, doi: 10.1002/sim.4053. PMID: 20809482.

Ghadessi, M., Tang, R., & Zhou, J. (2020). A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). *Orphanet J Rare Dis*, https://doi.org/10.1186/s13023-020-1332-x.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-Tests. *The American Statistician*, DOI:10.1080/00031305.2018.1562983.

Kelter, R. (2020). Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to Bayesian inference with JASP. *BMC Medical Research Methodology*, https://doi.org/10.1186/s12874-020-00980-6.

Li, W., Chen, M., & Wang, X. (2018). Bayesian Design of Non-inferiority Clinical Trials Via the Bayes Factor. *Stat Biosci* , https://doi.org/10.1007/s12561-017-9200-5.

Ly, A., Stefan, A., van Doom, J., Dablander, F., van den Bergh, D., Sarafoglou, A., . . . Wagenmakers, E.-J. (2020). The Bayesian Methodology of Sir Harold Jeffreys as a Practical Alternative to the P Value Hypothesis Test. *Computational Brain and Behavior*, https://doi.org/10.1007/s42113-019-00070-x.

Postma, D., van Werkhoven, C., van Elden, L., Thijsen, S., Hoepelman, A., Kluytmans, J., . . . Bonten, M. (2015). CAP-START Study Group. Antibiotic treatment strategies for community-acquired pneumonia in adults. *N Engl J Med*, doi: 10.1056/NEJMoa1406330. PMID: 25830421.

Robert, C. P. (2007). *The Bayesian Choice.* New York: Springer.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences. *Psychological Methods*, https://doi.org/10.1037/met0000061.

Schumi, J., & Wittes, J. T. (2011). Through the looking glass: understanding non-inferiority. *Trials*, https://doi.org/10.1186/1745-6215-12-106.

Stefan, A. M., Gronau, Q. F., Schonbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, https://doi.org/10.3758/s13428-018-01189-8.

Stefan, A. M., Katsimpokis, D., Gronau, Q. F., & Wagenmakers, E.-J. (2022). Expert agreement in prior elicitation adn its effects on Bayesian inference. *Psychonomic Bulletin and Review*, https://doi.org/10.3758/s13423-022-02074-4.

van Ravenzwaaij, D., Monden, R., Tendeiro, J., & Ioannidis, J. P. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Med Res Methodol*, https://doi.org/10.1186/s12874-019-0699-7.