Non-Inferiority Testing

Joel M Miller, PhD
jmm@eidactics.com

Eidactics • eidactics.com

Why?

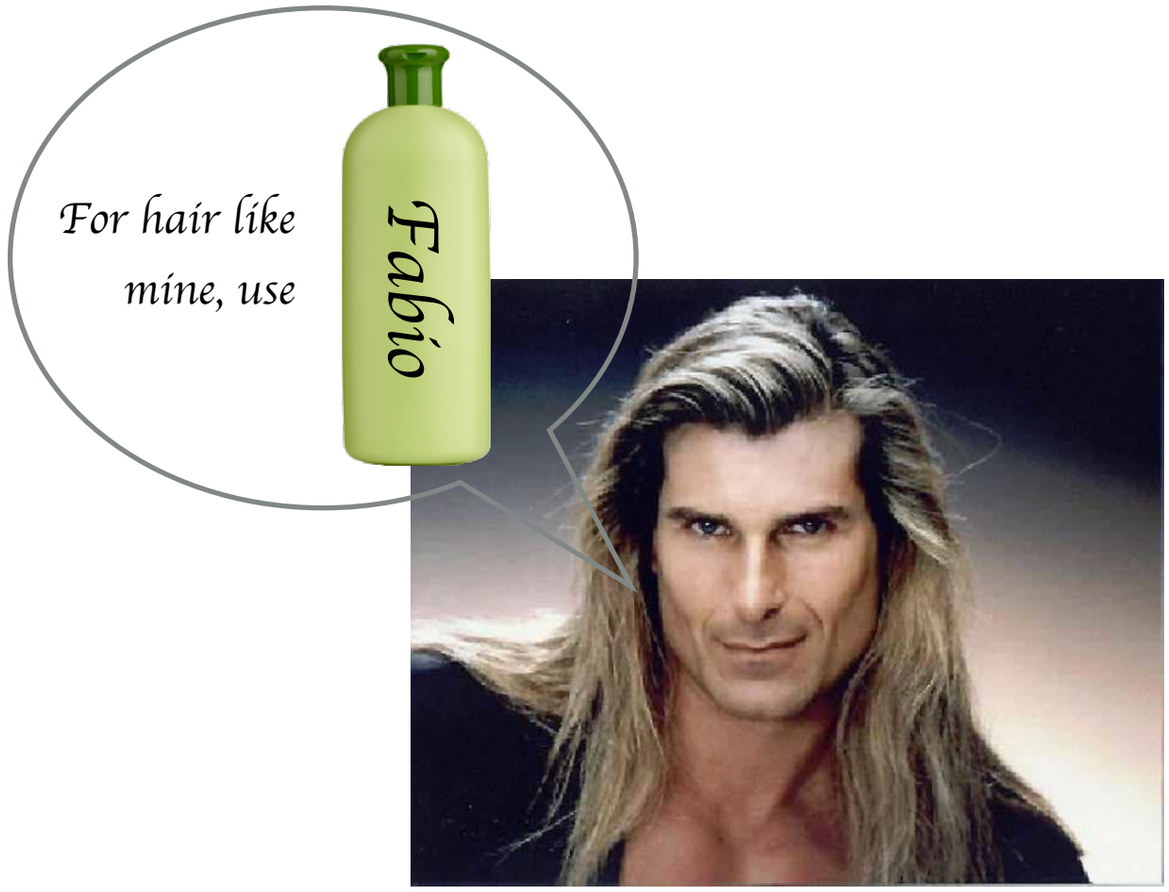
The conventional treatment for hair loss is Minoxidil. Suppose we've got a new treatment called "Fabio".

Ideally, we'd like to show Fabio is <u>better</u> than Minoxidil, so we get 100 pairs of balding men, and randomly give one of each pair Minoxidil, and the other Fabio.

We then measure hair growth, and for each pair of men compute the difference Fabio - Minoxidil.

Suppose we find F - M = 0.6.

Can we conclude that Fabio is better?



Is The Difference Real?

No, we cannot.

The reason is we know that if we repeated this experiment with another 100 pairs of men we'd get different results due to random variations (eg: Mean F-M = 0.5, 1.5, -0.3, 1.1, -1.5, …), and without further analysis we don't know whether, eg, F-M = 0.6 means Fabio really is better, or was merely a consequence of random variation.

How can we tell the difference?

That's what what statistical tests are for.

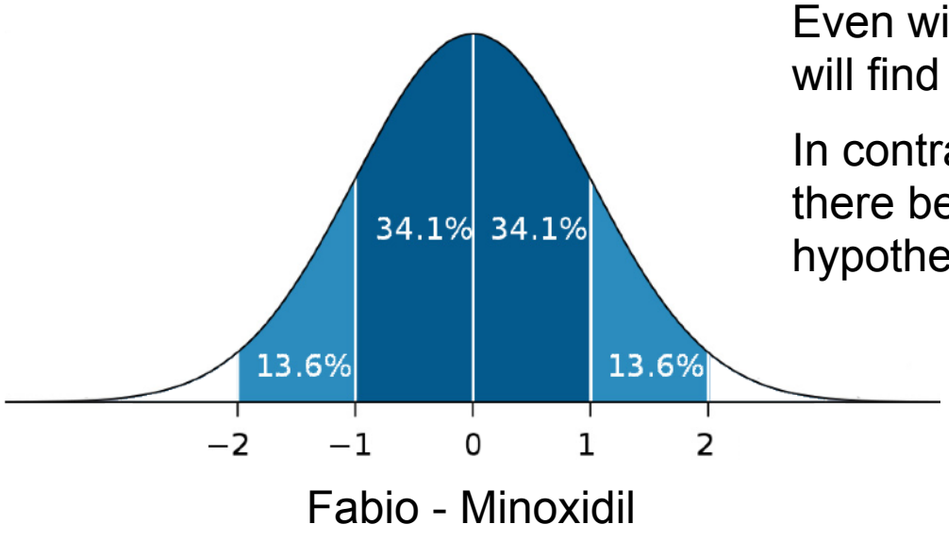
Superiority Testing

We can predict how such experiments are likely to turn out if we suppose we know the true F-M difference (and a few other things).

But, of course, the true F-M difference is exactly what we want to find, so we turn the question around:

- We assume there is no F-M difference – this is the null hypothesis.
- We estimate the probabilities of various possible outcomes of our experiment under that assumption – this prediction is called the sampling distribution.
- We see from the sampling distribution whether our actual experimental result was likely to to occur under the null hypothesis, that is, with there being no real F-M difference.

If our experimental result was unlikely to occur by chance under the null hypothesis, we say we have "rejected the null hypothesis", and conclude that there is a real F-M difference.



Even with no real difference between Fabio and Minoxidil, experiments like ours will find Mean F-M differences of 0 to 2 almost half the time, simply by chance.

In contrast, an experiment giving Mean F-M > 2 is very unlikely to occur with there being no real difference (only 2.5% probability), allowing us to reject the null hypothesis and to conclude that Fabio is better!

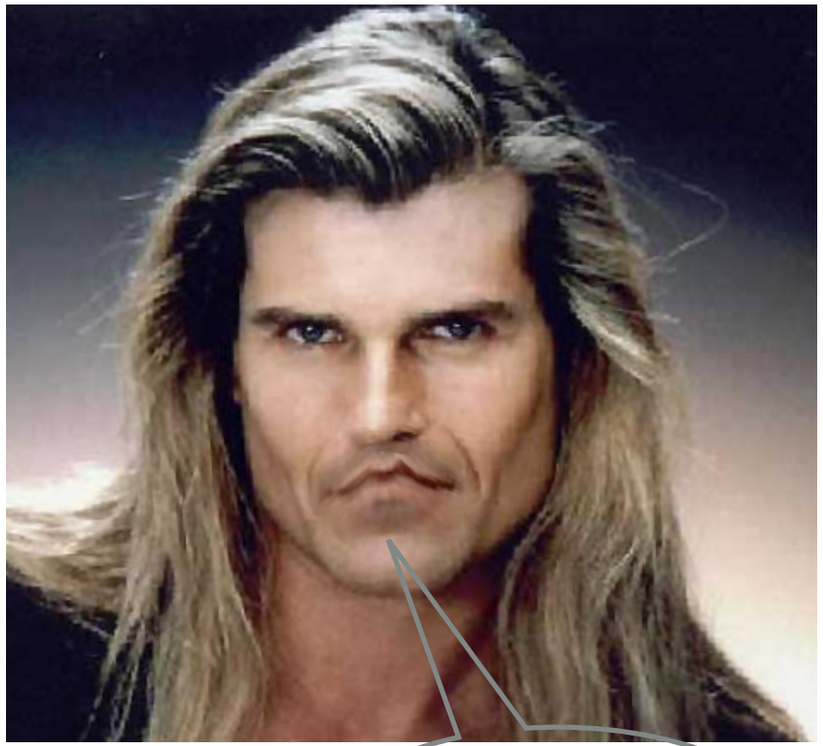
Are They The Same?

So, given our finding of F - M = 0.6 we cannot reject the null hypothesis of no difference. A difference of this size could have occurred by chance. Fabio, sadly, was not found to be significantly better than Minoxidil.

But what if we'd be satisfied with a weaker conclusion?

More generally, suppose it is unethical to test a new treatment against a placebo, or that we have a product or treatment that's superior on dimensions other than effectiveness (cost, ease of use, risk, . . .), and we want only to demonstrate that it is <u>equal in effectiveness</u> to an existing product or treatment.

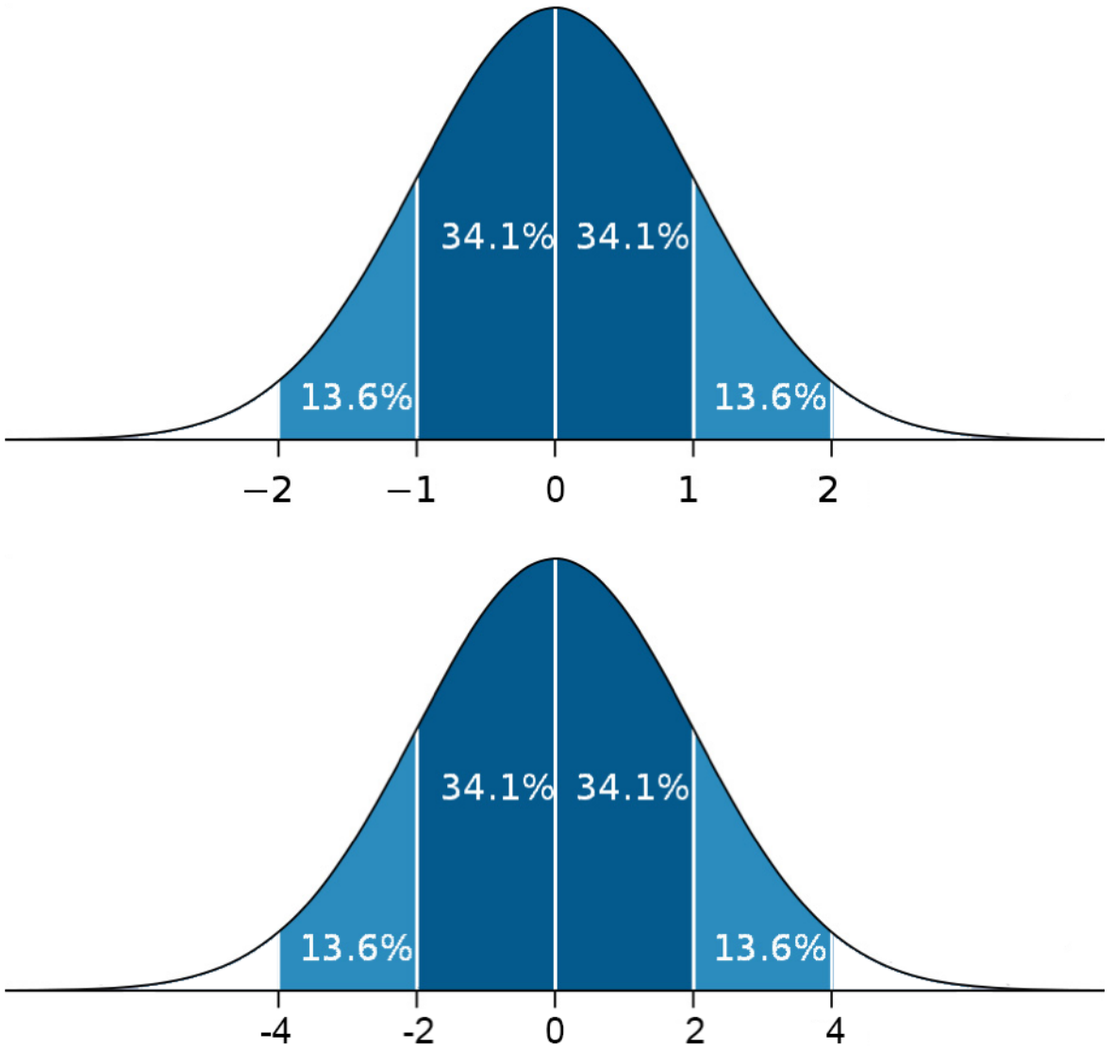
There is an obvious way to do this . . . and it is invalid.



Asserting The Null Hypothesis 1

If we fail to reject the null hypothesis, does that mean we can accept it and conclude that Fabio and Minoxidil are equal in effectiveness?

No it does not, because the sampling distribution gets wider as variability of the data increases and as the number of samples (pairs of men) in the study decreases:



In this well-controlled study of 100 pairs of men, a finding of Mean F-M > 2 would reject the null hypothesis.

But, with poorer controls, or fewer subjects, more variability is expected in experimental results, and Mean F-M > 2 might be expected 16% of the time, which is not so unlikely as to suggest a real difference.

Asserting The Null Hypothesis 2

Another way to look at it is that statistical tests are basically quotients:

Statistic = Difference of Group Means / Measure of Variability

If the statistic is greater that some prescribed value, the difference between groups is significant, otherwise not.

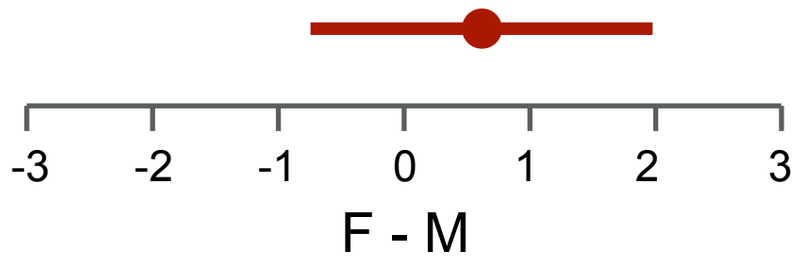
This makes sense:

- a big difference in group means (the numerator) is more likely to reflect a real difference between experimental groups, and less likely to be a result of random variation.
- a given difference in group means is more likely to reflect a real difference between groups, rather than random variability, if experimental variability (the denominator) is small.

Statistics and their significance depend on experimental variability, so failure to reject the null hypothesis may simply mean that you did a weak experiment (too much uncontrolled variability or not enough subjects), unable to detect a small, but real, difference.

Confidence Intervals

Suppose we do an experiment and find F - M = 0.6 (●). Due to random variation, of course we might have gotten a different result by chance.



Our data tells us something about experimental variability, and from it we can estimate the sampling distribution – what we would expect if we repeated our experiment.

From the sampling distribution we can say how likely it is that repeating our experiment would yield various results (F-M = 0.5, 1.5, -0.3, etc).

Let's consider the range we'd expect F-M results to fall in 95% of the time (▬). This is the 95% confidence interval, and we'll consider any result in this range to be just random variation, a result our experiment may well have yielded.

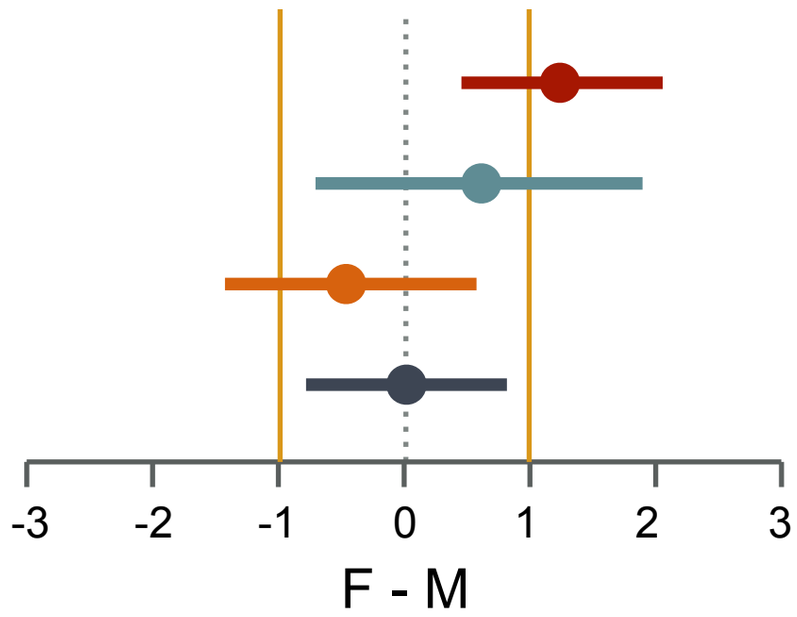
Conversely, it is very unlikely (<5% probability) that we could repeat our experiment and get a value outside this range (eg, F-M = 2.1 or -0.5). We agree to take such a result to indicate a real difference, not just random variation.

Superiority, Equivalence & Noninferiority

Practically, no two treatments are exactly equivalent, so what equivalence testing really determines is whether two treatment effects are "close enough", that is, within an equivalence margin.

In the figure, "0" means Fabio & Minoxidil have exactly the same effect, and "-1" means Fabio is one unit worse.

Suppose we set the equivalence margin at ±1 unit, a range we feel defines practically equivalence (yellow line).

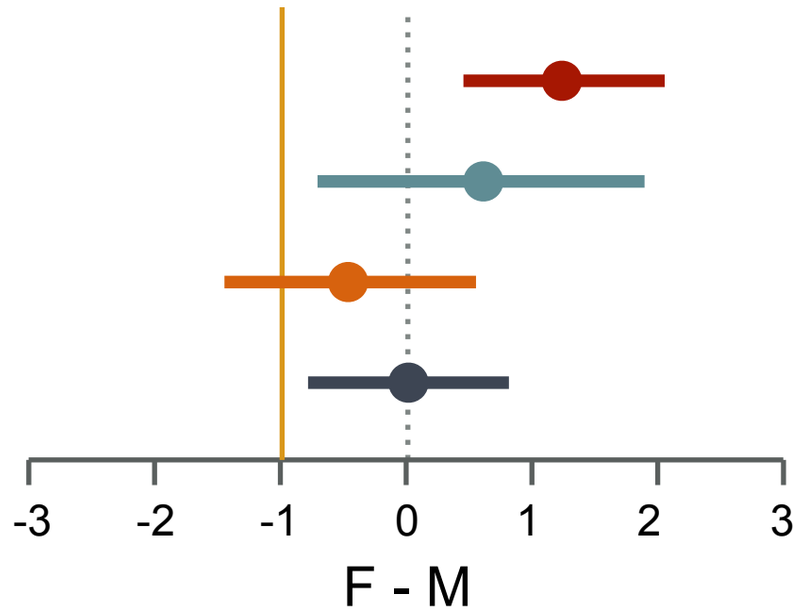


- Fabio is superior to Minoxidil – a conventional difference test demonstrates that F-M is significantly different from zero.
- Fabio is not shown to be superior, but is shown to be noninferior to Minoxidil (the "significantly worse" line at -1 falls outside the confidence limit).
- Fabio is not shown to be noninferior (a "significantly worse" outcome might well be found if this experiment were repeated).
- Fabio is equivalent to Minoxidil – it is unlikely to be significantly better or worse.

Noninferiority

Noninferiority tests aim to show that an experimental treatment is not less effective than an active control (eg, an existing treatment) by more than the equivalence margin.

A noninferiority design tests the null hypothesis that the experimental treatment is inferior by more than the equivalence margin, and seeks to reject that hypothesis.

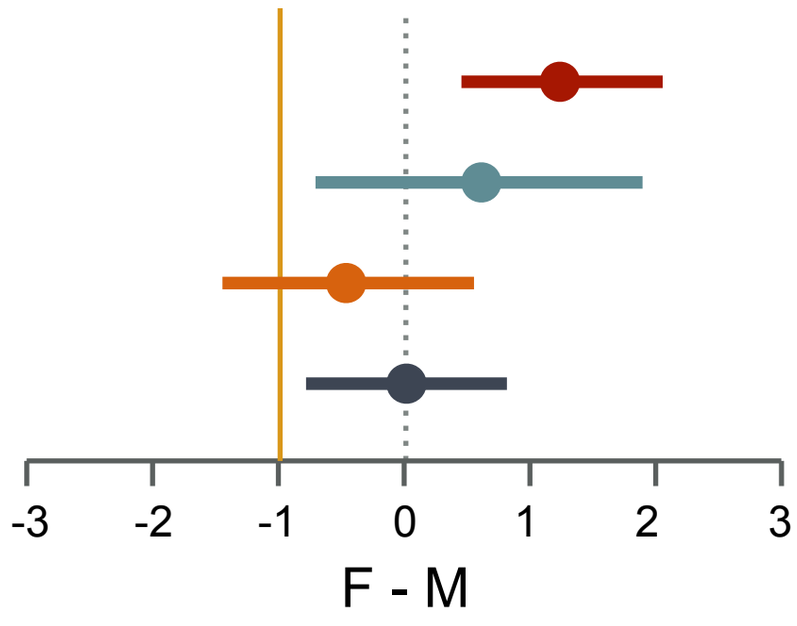


- Fabio is noninferior – its effectiveness is unlikely to fall below the lower equivalence margin on further trials (in fact, we can make the stronger claim that it is superior).
- Fabio is noninferior (though not superior).
- Fabio is not shown to be noninferior.
- Fabio is noninferior, (and also non-superior, so we say it is equivalent).

Multiple Comparisons

If you have multiple groups (given, say, 8 different drugs), it is obviously invalid to make all 28 possible comparisons and then report as "significant differences" those paired comparisons that differ at $p < 0.5$.

There may be a similar problem performing both superiority testing and noninferiority testing on the same data. It seems safest to decide in advance which type of testing one is doing and stick to that decision (right!).



It may be acceptable, however, to plan a noninferiority trial, but then report superiority if a result like —●— is found, where the confidence interval excludes not only the noninferiority margin but also zero.

But, if a superiority trial fails to reject the null hypothesis, it is considered invalid to "fall back" and report noninferiority without some appropriate statistical adjustment.

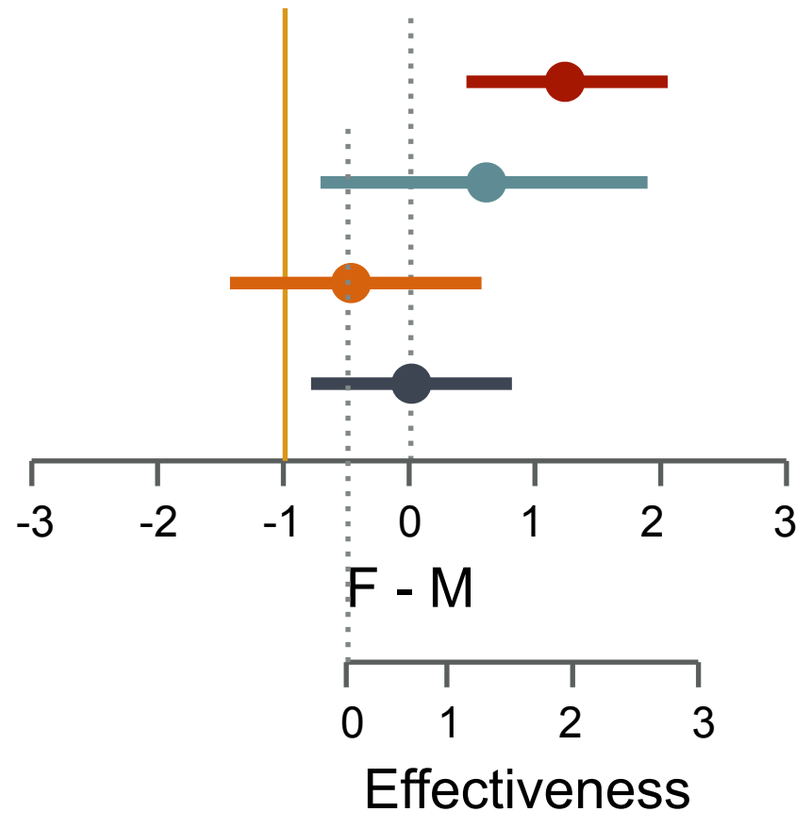
Does Noninferiority Imply Effectiveness?

Suppose Minoxidil is known to be effective, specifically to grow an average of 0.8 units of hair.

If Fabio is shown to be noninferior to Minoxidil, does that imply it too is effective?

Sadly, it does not!

Noninferior to an effective treatment, but nevertheless, not significantly effective.



Thus:

- Equivalence margins must be chosen carefully.
- Testing against a placebo, as well as against the established treatment may be useful.