The new bee guidance draft
Using equivalence tests

Ines Hotopp
Markus Persigehl
Andrea Rossbach

Introduction

The guidance draft for the risk assessment of plant protection products on bees asks for the use of equivalence instead of difference tests. An equivalence tests is based on the assumption that there is a risk as null hypothesis, and thus places the burden to proof a low risk on the opposite side. A 10% deviation from the baseline is deemed to be the threshold to distinguish between a low and a high risk. The description given in the draft leaves open the exact definition of this baseline and implicitly of the acceptable effect size delta. There are three options and the choice is crucial for the results of the statistical analysis.

Here, we compare equivalence and difference testing methods using the data from 48 control honeybee colonies of a field effect study conducted in northern Germany in 2014 (Rolke et al. 2014) and applying two of the options for the baseline definition.

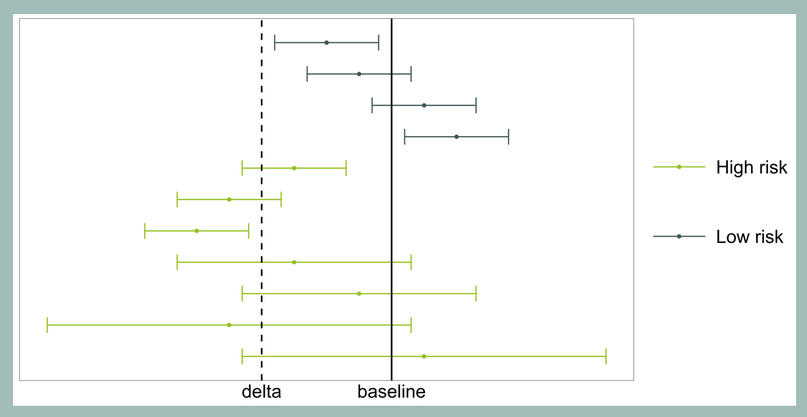
Definition of the baseline

- For an adverse effect on the colony strength

Option	Baseline	delta
A	Control mean	Control mean - 10% of control mean
B	Lower CL of control mean	Lower CL of control mean - 10% of control mean

- Option A
  - True extend of natural variability of control not included
  - Only 10% variability allowed
- Option B
  - Extend of natural variability of control included
  - Additional 10% effect allowed

Definition of low and high risk



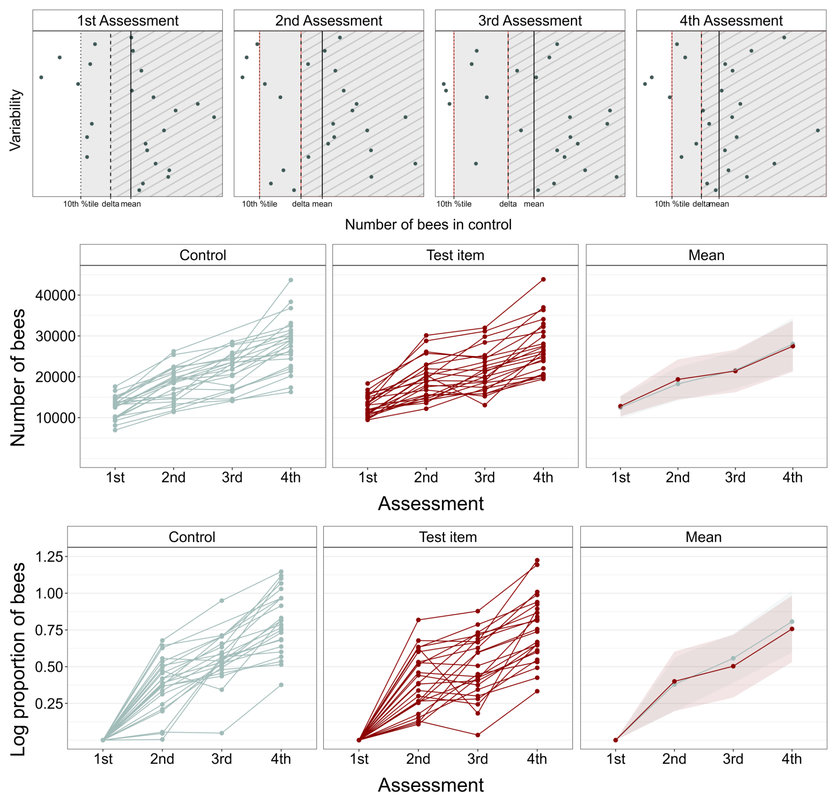
Abbreviations:
CL Confidence limit
SD Standard deviation
GLMM Generalized linear mixed model

Method

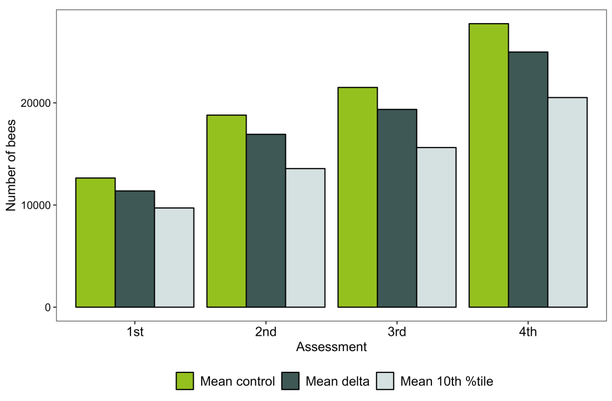
- Honeybee field effect study in Northern Germany with 48 control colonies with initially about 13,000 adult bees on 6 fields
- Estimation of the colony strength in 4 weekly colony assessments
- In 10 x 1 000 runs 24 colonies randomly selected as hypothetical test item colonies
- Assessment of variability
  - Control mean
  - Delta: control mean – effect size x control mean
  - 10^{th} percentile of control (90% of colonies above this value)
- Assessment-wise statistical analysis of the colony strength
  1. GLMM as difference tests
  2. GLMM as equivalence tests using the control mean as baseline
  3. GLMM as equivalence tests using the lower control confidence limit as baseline
  4. Welch's t-test as difference tests
  5. Welch's t-test as equivalence tests
- GLMM model with negative binomial family and formula

  Colony strength ~ Treatment x Assessment + (1|Field/Colony)
- t-test calculation with log-transformed proportion of the colony strength relating to the first assessment
- Difference tests with alpha = 0.05 using R-functions *glmmMB()* and *t.test()*
- Equivalence tests with alpha = 0.2 and effect size = 10% using R-functions *glmmTMB()*, *emmeans()* and *tsum_TOST()*
- Calculation of the probability to find a significant effect for difference tests
- Calculation of the probability for not being able to exclude a high risk for equivalence tests

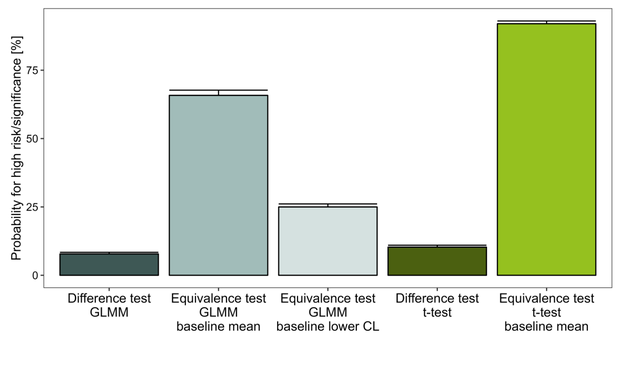
Example combination



Results



Variability between control colonies was higher than 10%. The difference between control mean and the 10^{th} %tile increased from 23 to 26% from the 1^{st} to the 4^{th} assessment. The probability for the detection of a significant effect with difference tests was 8% for a GLMM and 10% for a t-test. The probability to not be able to exclude a high risk was 92% for equivalence tests conducted with a t-test. For GLMMs it was 66% when using the mean of the control as baseline for the definition of delta. When using the lower CI, this probability dropped to 25%.



Conclusion

The data used in this analysis had - for a honeybee field effect study - an unusual high number of colonies. The colonies started with a similar number of bees and great effort was undertaken to achieve a similar development. Still, the variability of the colony strength expressed as the 10% tile was more than twice as high as the 10% effect size set by EFSA with an increase throughout the study. Following this observation, equivalence tests based on the control mean were able to exclude a high risk with only a unsatisfactory probability. However, when using the lower CI of the control and thus including the natural variability of the control colonies, resulted –at least for GLMMs – is an acceptable probability to prove that no high risk is to be expected.

tier3 solutions GmbH,
Kolberger Str. 61-63,
Leverkusen, Germany,
e-mail:
ines.hotopp@tier3.de