# Assessing the borderline range of prediction models: Method and implications for decision-making

Silke Gabbert[1]

Susanne N. Kolle, Miriam Mathea, Bennard van Ravenzwaay, Robert Landsiedel[2]

[1]Wageningen University & Research, The Netherlands

[2]BASF SE, Germany

BASF
We create chemistry

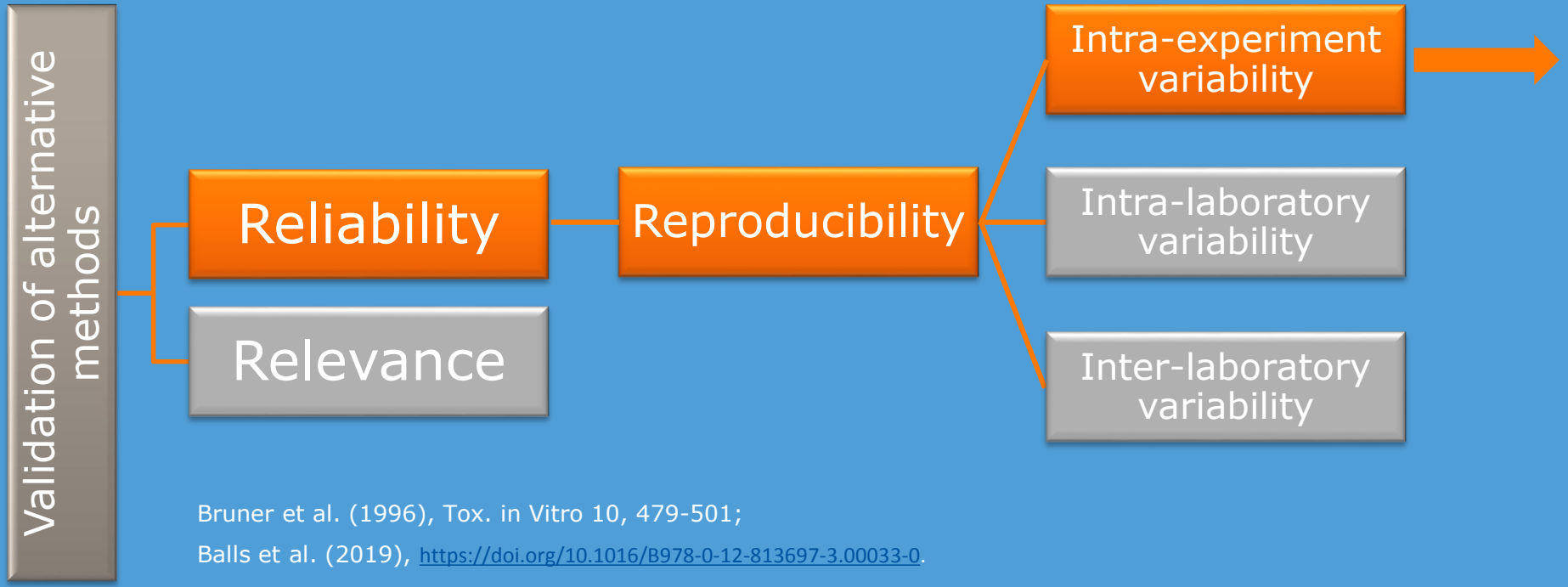
WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

# Outline

1. Motivation
2. Methods for quantifying the borderline range
3. Test selection and datasets
4. Results
5. Discussion
6. Conclusions and outlook

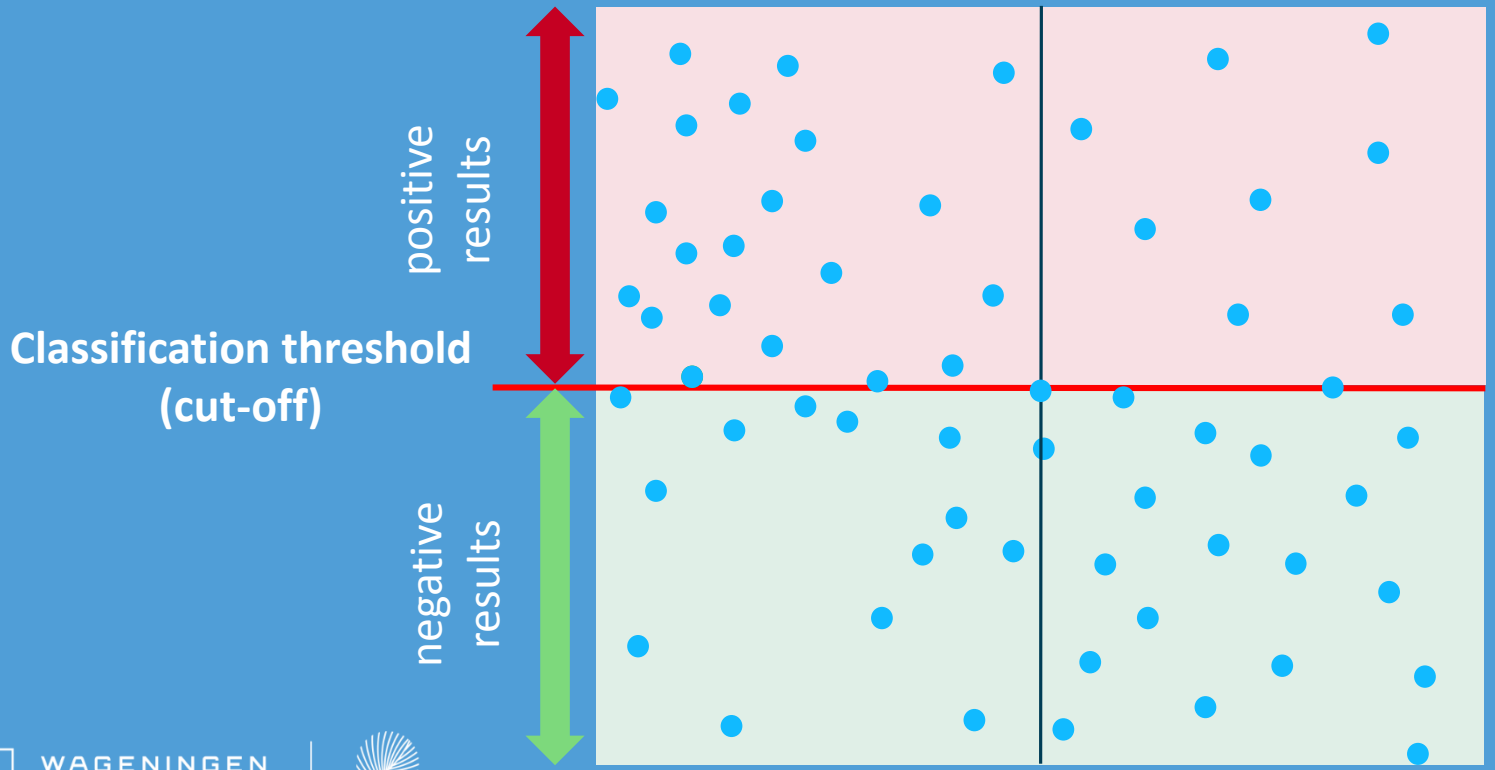
# 1. Motivation



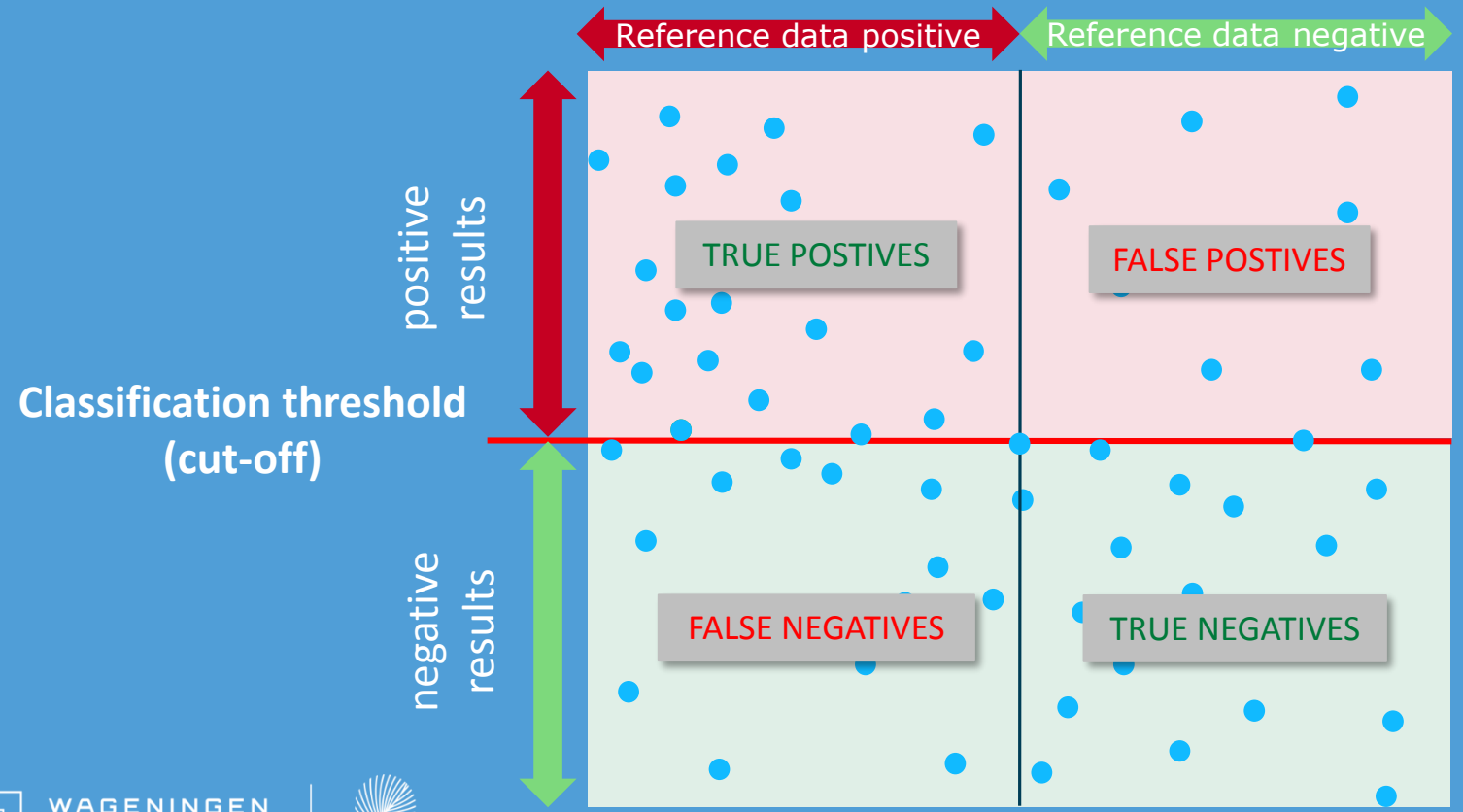

Bruner et al. (1996), Tox. in Vitro 10, 479-501;

Balls et al. (2019), https://doi.org/10.1016/B978-0-12-813697-3.00033-0.

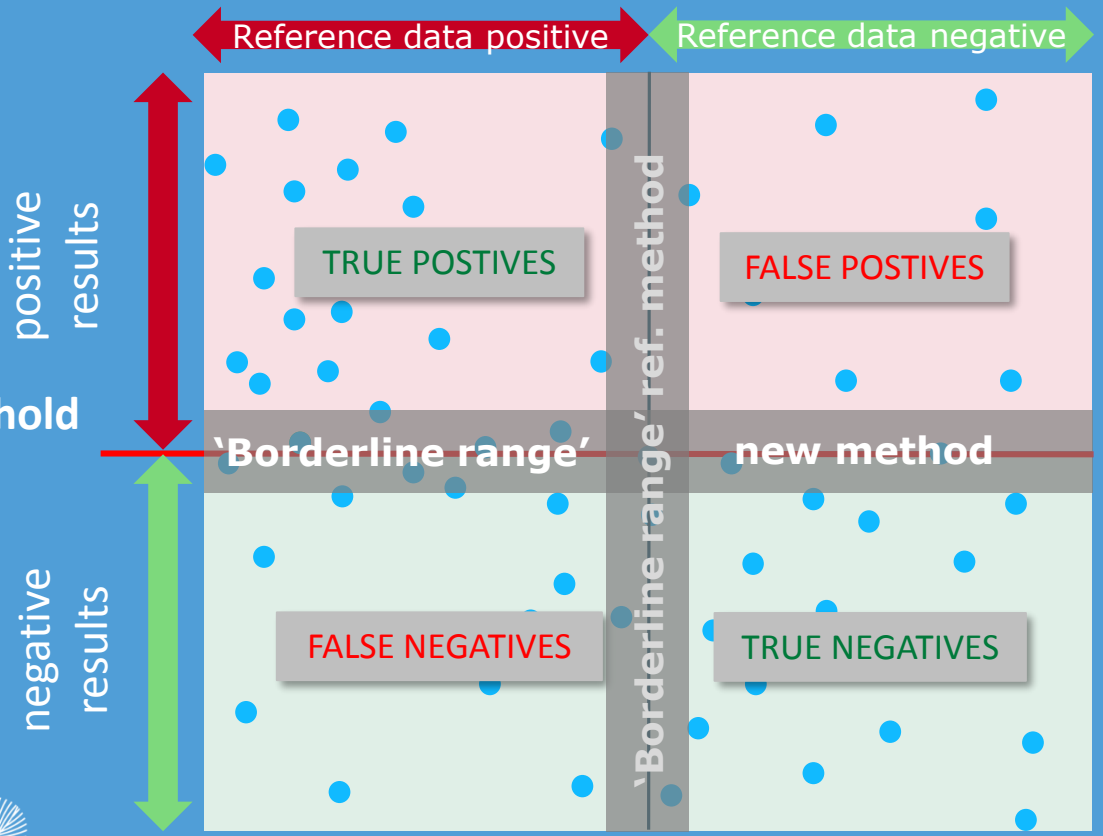
# 1. Motivation



# 1. Motivation



# 1. Motivation



# 1. Motivation

**Definition of the borderline range:**

Range around the classification threshold where the probability to obtain a positive and a negative result are equal

→ Results in the BR cannot unambiguously be classified!

**Key questions:**

1. How can the borderline range be determined?
2. Does the borderline range vary across alternative methods?
3. What are the implications for assessing alternative methods' predictive capacity?

# 2. Methods for quantifying in borderline range

## Previous research:

| Assay | Endpoint | Cut-off | TG borderline range | Dataset size | BR (based on pooled SD) |
|---|---|---|---|---|---|
| **DPRA** (OECD TG 442C) | Mean peptide depletion [%] | 6.38 | 3-10 | 42 | 4.86-7.90 |
| | Cysteine-only depletion [%] | 13.89 | 9-17 | 42 | 10.50-17.28 |
| **LuSens** (OECD TG 442D) | Luciferase fold induction | 1.50 | n/a | 26 | 1.27-1.73 |
| **h-CLAT** (OECD TG 442E) | CD54 induction | 200 | n/a | 13 | 181-219 |
| | CD86 induction | 150 | n/a | 13 | 124- 176 |
| **LLNA** (OECD TG 429) | Thymidine incorporation | 3.0 | n/a | 22 | 2.20- 3.71 |

Leontaridou et al. (2017): ALTEX 34(4), 525-538.

# 2. Methods for quantifying the borderline range

## This study: Comparison of methods

| Method for calculating the BR | BR |
|---|---|
| 1. Pooled standard deviation | $BR = \{T - SD_p \,; T + SD_p\}$ |
| 2. Pooled median absolute deviation (MAD) | $BR = \{T - MAD_p \,; T + MAD_p\}$ |
| 3. Confidence interval approx. of 2. using bootstrap percentile method | $BR = \{T - CI_{0.9} \,; T + CI_{0.9}\}$ |
| 4. 90% percentile of the distribution of all MADs | $BR = \{T - MAD_{0.9} \,; T + MAD_{0.9}\}$ |

Gabbert et al. (2019), in progress.

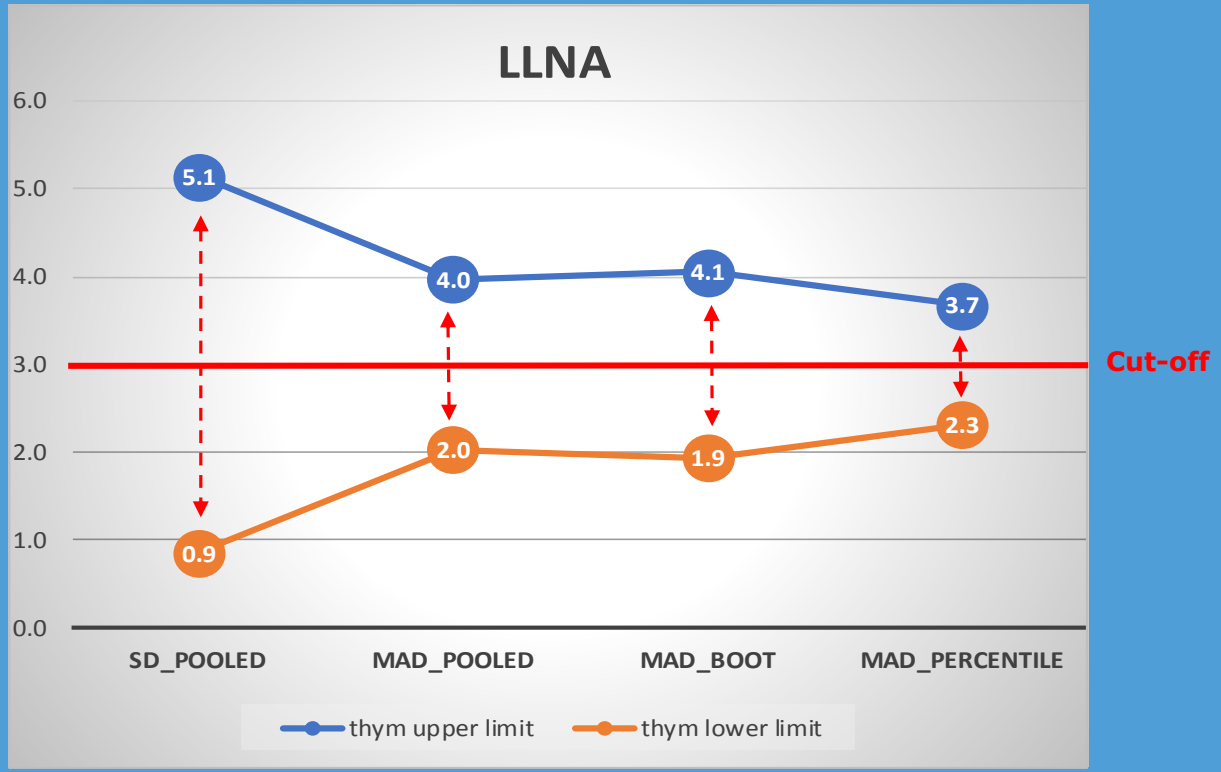
# 3. Test selection and datasets

## This study: Data

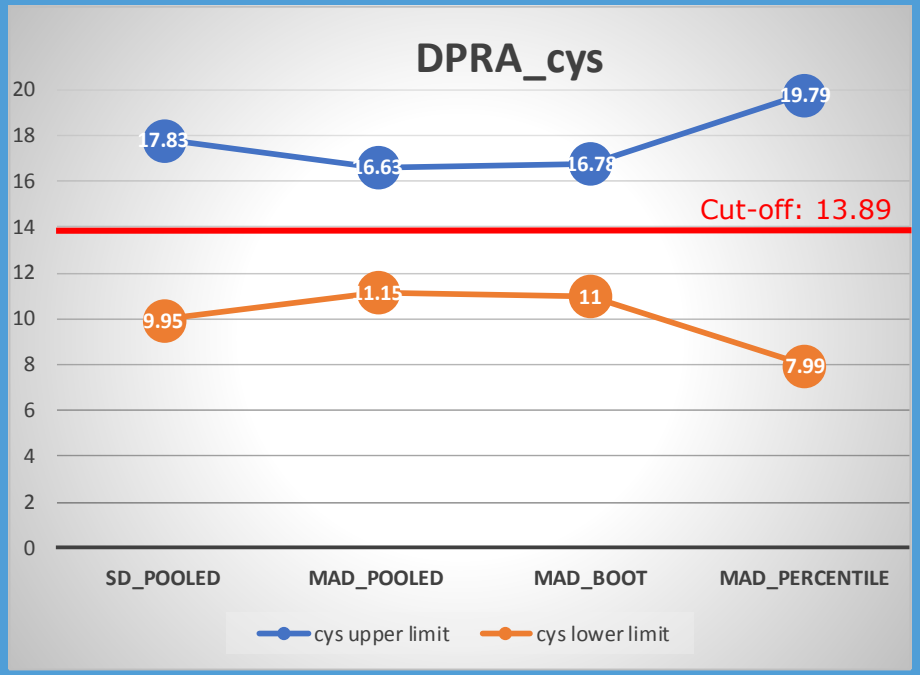
| Assay | Endpoint | Cut-off | TG-BR | Data set size | BR | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Method 1 pooled SD | Method 2 MAD | Method 3 bootstrap | Method 4 90% percent. |
| **DPRA** (OECD TG 442C) | Mean peptide depletion [%] | 6.38 | 3-10 | 138 | 4.1-8.4 | 4.8-8.0 | 4.7-8.1 | 2.7-10.1 |
| | Cysteine-only depletion [%] | 13.89 | 9-17 | 138 | 10.0-17.8 | 11.2-16.6 | 11.0-16.8 | 8-19.8 |
| **LuSens** (OECD TG 442D) | Luciferase fold induction | 1.5 | n/a | 130 | 1.2-1.8 | 1.4-1.6 | 1.4-1.6 | 1.3-1.7 |
| **h-CLAT** (OECD TG 442E) | CD54 induction | 200 | n/a | 134 | -0.6-401 | 125-275 | 121-279 | 51-349 |
| | CD86 induction | 150 | n/a | 134 | 173-127 | 135-165 | 135-165 | 117-183 |
| **LLNA** (OECD TG 429) | Thymidine incorporation | 3 | n/a | 68 | 0.9-5 | 2-4 | 2-4 | 2-4 |

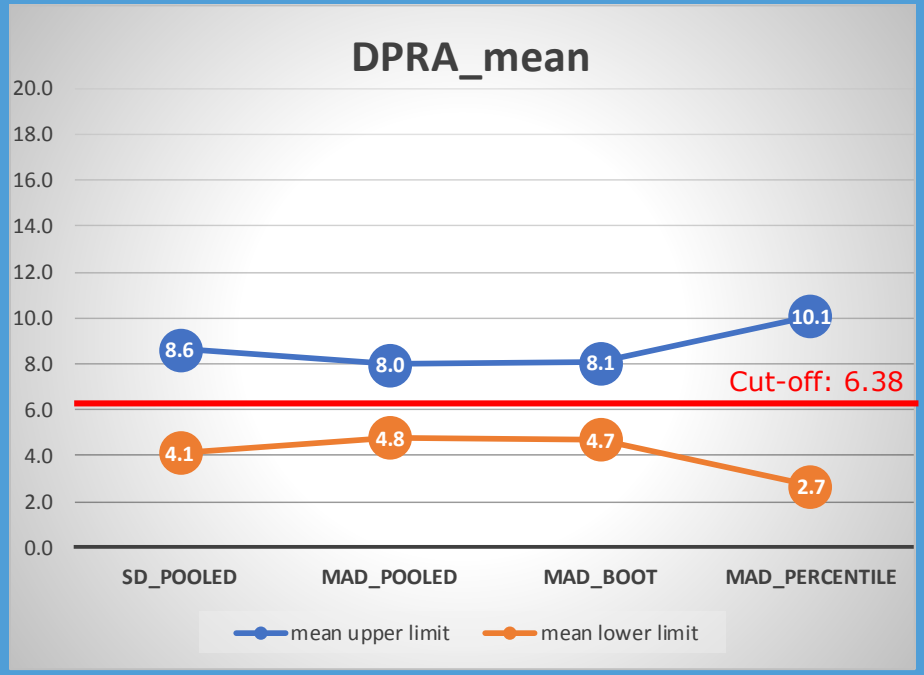
Gabbert et al. (2019), in progress.

# 4. Results - LLNA

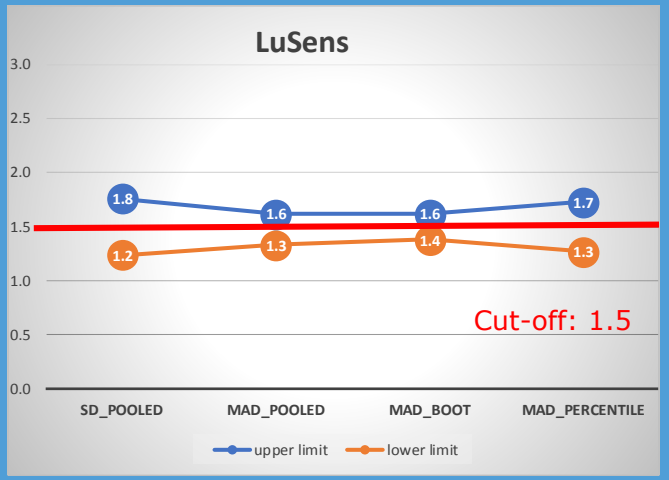


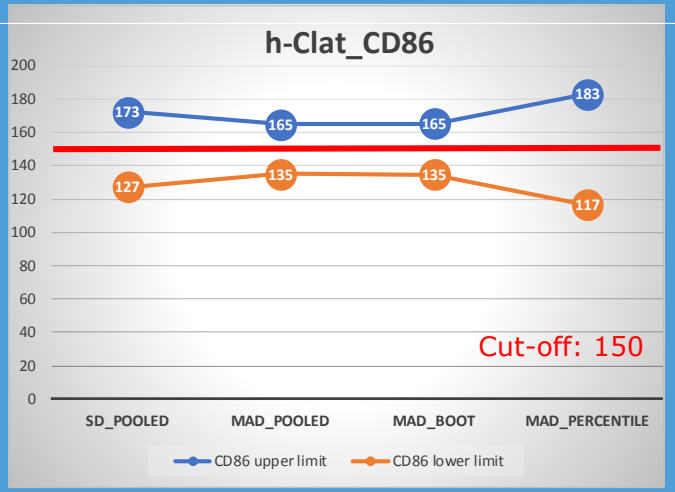
# 4. Results - DPRA

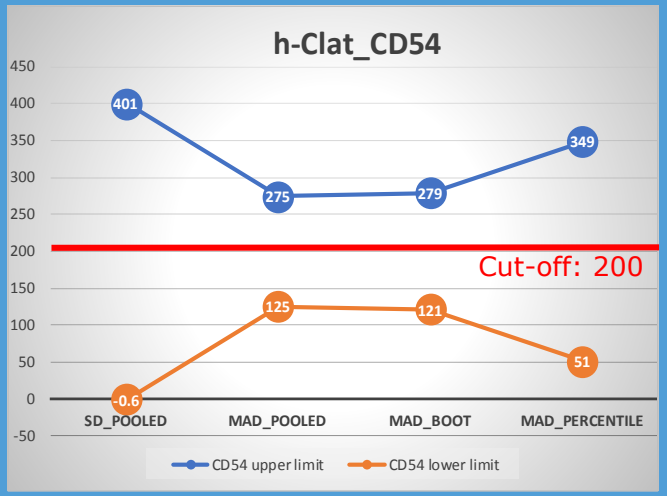


# 4. Results – LuSens and h-Clat



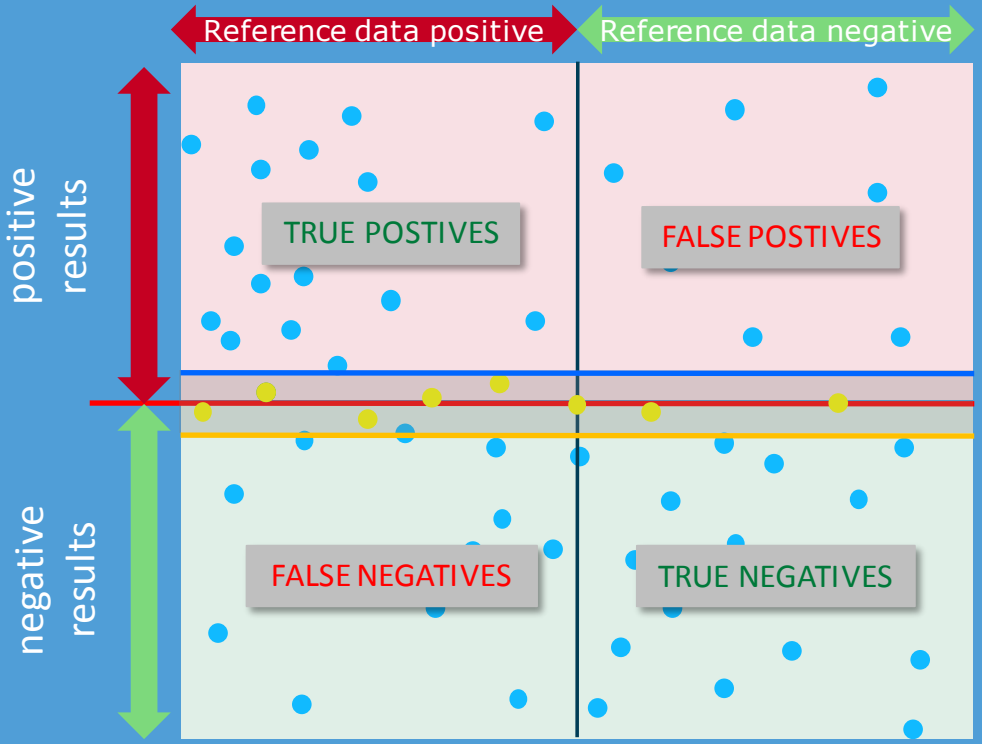




WAGENINGEN
UNIVERSITY & RESEARCH

# 4. Results

- The size of the BR differs depending on the statistical approach
- The BR is smallest when calculated as MAD (including bootstrap variant)
- Irrespective of how the BR is determined it has implications for determining a (non-animal) test method's predictive accuracy

# 5. Discussion




$$\text{Sensitivity [\%]} = \frac{TP}{TN+FN} * 100$$

$$\text{Specificity [\%]} = \frac{TN}{TN+FP} * 100$$

$$\text{Accuracy [\%]} = \frac{TP+TN}{TP+TN+FP+FN} * 100$$

# 5. Discussion

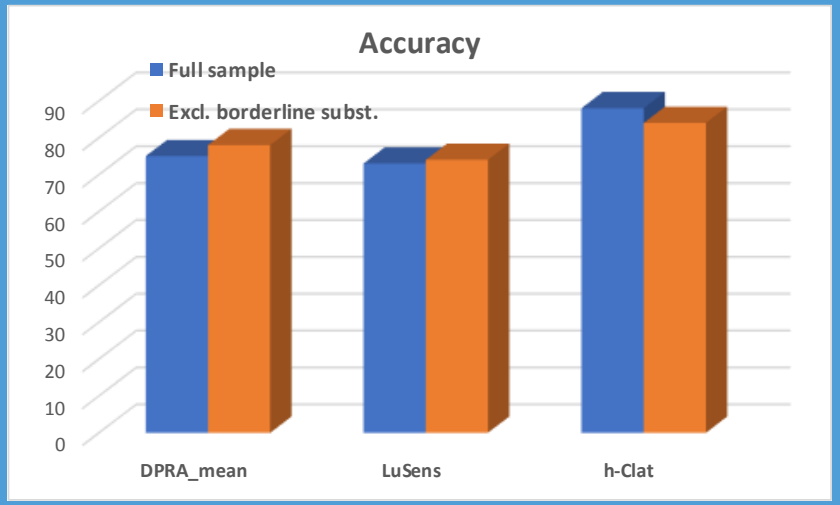
## Impact of the BR on predictive accuracy metrics:

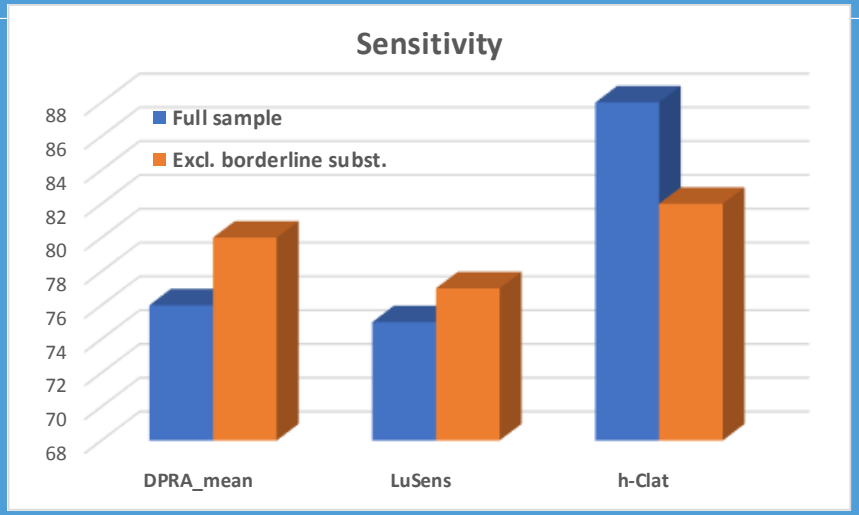
| Assay | Endpoint | Cut-off | Dataset incl. BR substances | Dataset excl. BR substances | Number and (%) of BR substances |
|---|---|---|---|---|---|
| **DPRA** (OECD TG 442C) | Mean peptide depletion [%] | 6.38 | 199 | 179 | 20 (10) |
| **LuSens** (OECD TG 442D) | Luciferase fold induction | 1.50 | 79 | 74 | 5 (6) |
| **h-CLAT** (OECD TG 442E) | CD54 induction | 200 | 40 | 32 | 8 (20) |
| | CD86 induction | 150 | 40 | 32 | 8 (20) |

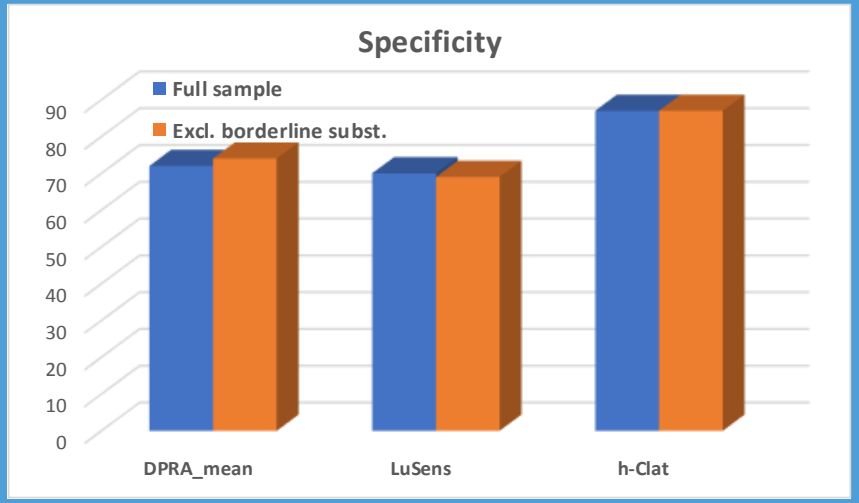
Leontaridou et al. (2019), under review.

# 5. Discussion

## Impact of the BR on predictive accuracy metrics:



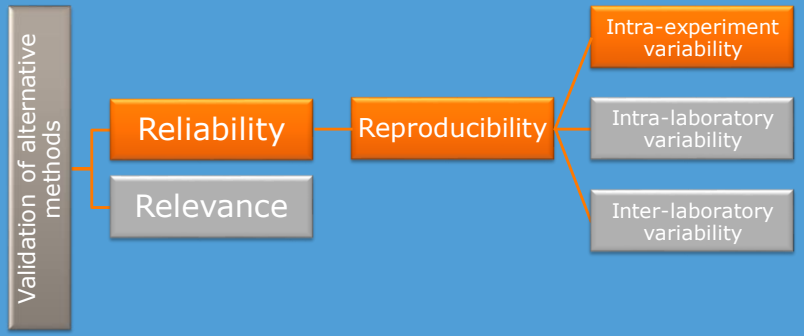




Leontaridou et al. (2019), under review.

# 6. Conclusions and outlook

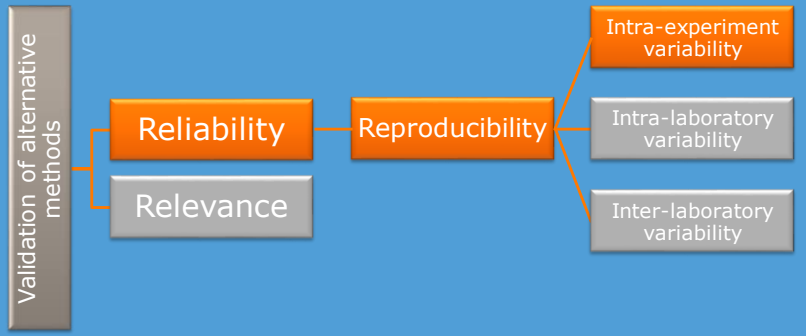
- Intra-experiment variability of test results was quantified as borderline range around a test's classification threshold.




- The BR impacts a (non-animal) method's reliability and, thus, its predictive accuracy.

- In the draft OECD guideline for Defined Approaches Skin Sensitization the BR has not been addressed.

# 6. Conclusions and outlook




- Further research should address
  -the interdependence of the BR in the animal and the non-animal test;
  -the combined impact of the BR, experimental sample size and sample composition on predictive accuracy;
  -how 'the optimal size' of the BR can be determined.

# Thank you!

Any comments welcome

WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

BASF
We create chemistry