Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Measuring calibration of LR-systems

Peter Vergeer, p.vergeer@nfi.nl
Netherlands Forensic Institute

Outline:

Why calibrate LR-systems?
- Calibration and Bayes decisions

Why measure calibration?
- Three reasons why
- Can one mess things up by being ill-calibrated?

Measuring calibration of LR-systems
- Visual representations
- Metrics

Calibration and Bayes decisions

Imagine a binary decision $\boldsymbol{D_j}$ with j = 1,2, based on costs for errors and probabilities for $\boldsymbol{H_i}$

Truth

Decision↓	$\boldsymbol{H_1}$	$\boldsymbol{H_2}$
$\boldsymbol{D_1}$	0	$\boldsymbol{C_{12}}$
$\boldsymbol{D_2}$	$\boldsymbol{C_{21}}$	0

Cost notation → $\boldsymbol{C_{DH}}$

Decision↓	H_1	H_2
D_1	0	C_{12}
D_2	C_{21}	0

Calibration and Bayes decisions: theory

Bayes decisions minimize perceived (or expected) costs:

D_1 if $C_{12}P(H_2|LR) < C_{21}P(H_1|LR)$

Else: D_2

With known priors the decision is based on the likelihood ratio:

D_1 if $LR > \frac{C_{12}}{C_{21}} \times \frac{P(H_2)}{P(H_1)} = LR_{th}$

Else: D_2

Decision↓	H_1	H_2
D_1	0	C_{12}
D_2	C_{21}	0

Calibration and Bayes decisions: practice

Does this minimize costs in reality?

- Variance? → Repeating the decision process reduces variance.
- Bias (leads to increased costs)?
  - Not when: $P(H_1|LR)$ and $P(H_2|LR)$ are well-calibrated!

Well-calibrated meaning: $freq(H_1|P(H_1|LR) = X) = X$, for all X

Focussing on LR-systems: $\frac{freq(LR=Y|H_1)}{freq(LR=Y|H_2)} = Y$, for all Y ('The LR of the LR is the LR')

Why measure calibration?

- Three reasons why:
  1. If we do not mean 'The LR of the LR is the LR', updating prior odds with Bayes rule would result in (very) misleading posterior odds
  2. We are not optimal from a Bayesian decision perspective
  3. We could do worse than not updating by Bayes rule

- Calibration should be measured in order to be sure to prevent the above

Can one mess things up by being ill-calibrated?

Compare using an LR-system to not using an LR-system

With LRs:

$\boldsymbol{D_1}$ if $\boldsymbol{C_{12}} \times \boldsymbol{P(H_2|LR)} < \boldsymbol{C_{21}} \times \boldsymbol{P(H_1|LR)}$
Else: $\boldsymbol{D_2}$

Only priors

$\boldsymbol{D_1}$ if $\boldsymbol{C_{12}} \times \boldsymbol{P(H_2)} < \boldsymbol{C_{21}} \times \boldsymbol{P(H_1)}$
Else: $\boldsymbol{D_2}$

We can compare the expected costs of the two…

Can one mess things up by being ill-calibrated?

LR-system for a binary observation:

$$LR(Obs"1") = 90$$
$$LR(Obs"2") = 0.101$$

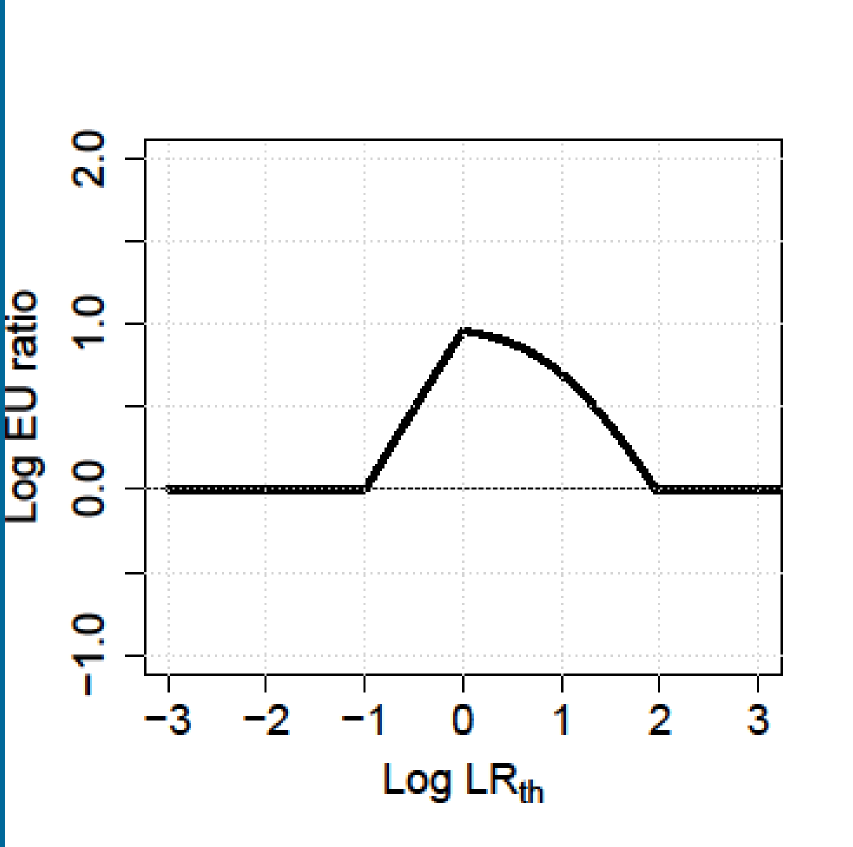
Experiment	H_1	H_2
Obs “1”	90	10
Obs “2”	10	990

Well-calibrated LRs according to experiment:

$$LR(Obs"1") = \frac{90}{100} * \frac{1000}{10} = 90$$
$$LR(Obs"2") = \frac{10}{100} * \frac{1000}{990} = 0.101$$

Can one mess things up by being ill-calibrated?

$$\frac{Empirical\ costs\ (LR = 1\ always)}{Empirical\ costs\ (LR - system)}$$



Brummer PhD thesis
Vergeer et al. S&J56(2016)482

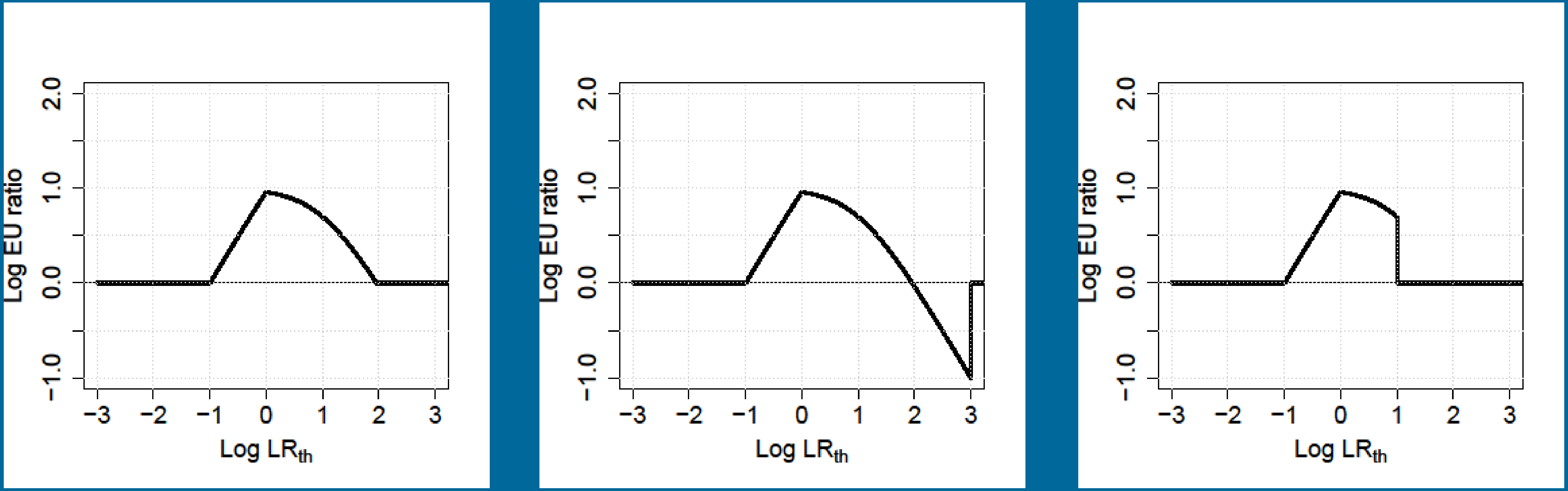
Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Can one mess things up by being ill-calibrated?

Well-calibrated
$LR(Obs"1") = 90$
$LR(Obs"2") = 0.101$

Overconfident
$LR(Obs"1") = \mathbf{1000}$
$LR(Obs"2") = 0.101$

Too conservative
$LR(Obs"1") = \mathbf{10}$
$LR(Obs"2") = 0.101$



Data from experiment remains the same

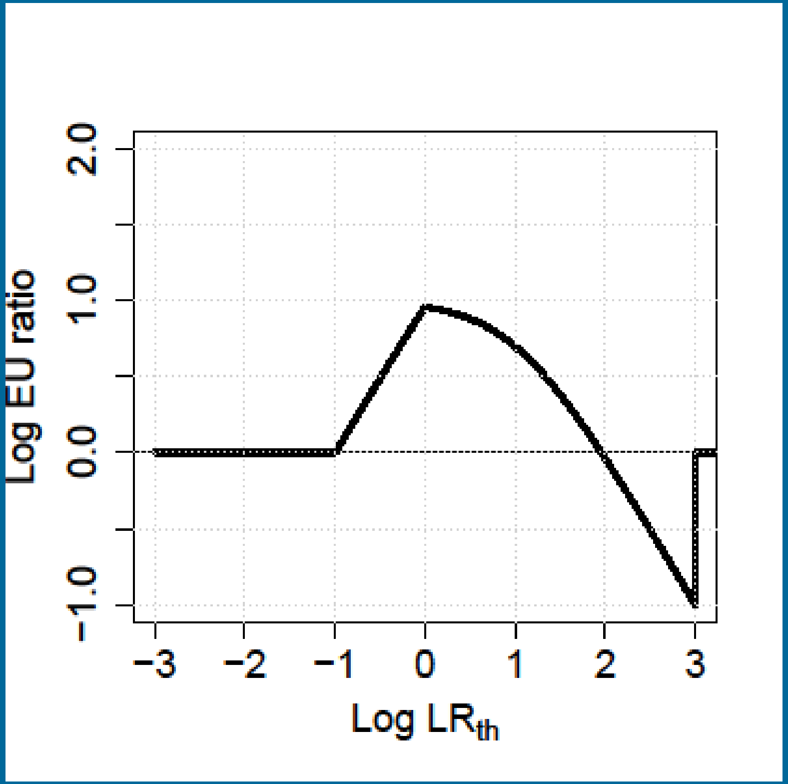
Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Can one mess things up by being ill-calibrated?

Overconfident

$LR(Obs"1") = 1000$

$LR(Obs"2") = 0.101$



In certain cost/prior scenarios using the LR-system results in worse performance!

Problem:
- Cost/prior scenario is variable.
- Cost/prior scenario is unknown.

Solution:
Measure calibration! → Well-calibrated LR-systems perform better than prior-only for all cost/prior scenario's.

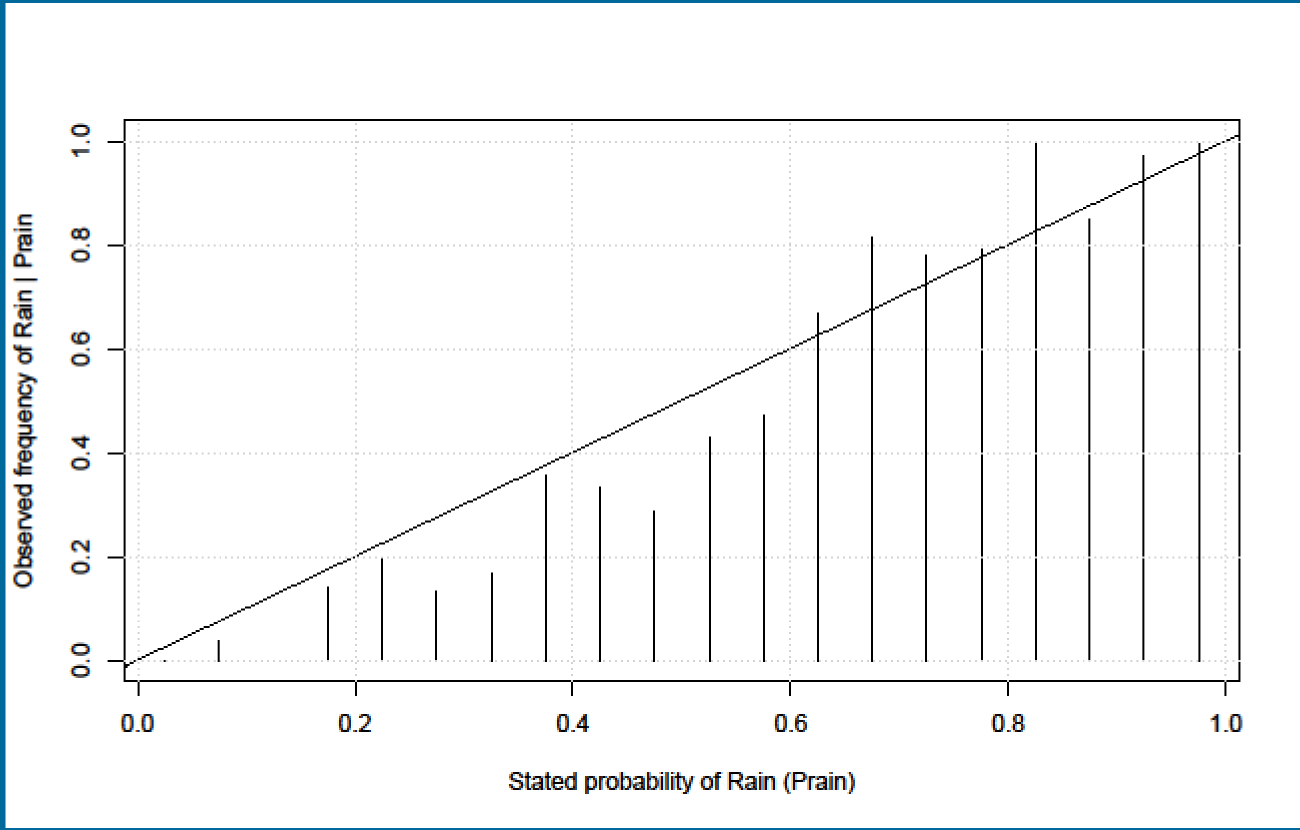
Measuring calibration of probabilities: a visual tool

Measuring calibration of probabilities

Collect stated P's for rain and the truth "rain" or "no rain" for a sequence of days.

Plot $freq(rain|Prain = X)$ versus X

'Binning and counting frequencies'



A 'calibration plot' for probabilities

Measuring calibration of LR-systems

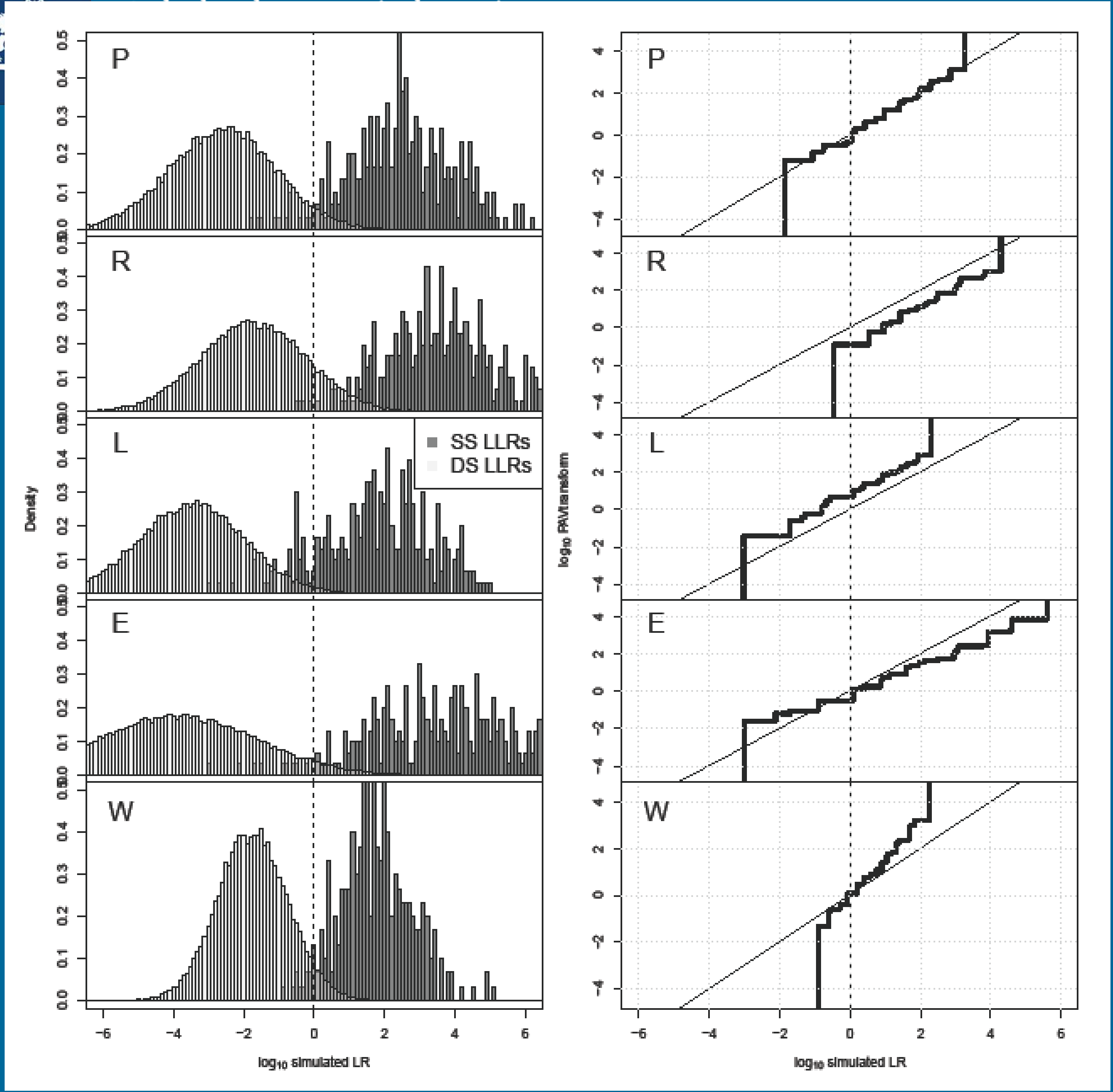
- Visual representations
  'LR of LR = LR' ?

- All based on some form of calculating observed LRs based on counts/frequencies.

- Main difference: automated binning or predetermined binning?

- PAV transform → automated binning with attractive theoretical properties

See e.g.:
Dawid JASA 77 (1982) 605
Ramos et al. FSI 230 (2013) 156
Vergeer et al. SciJus 57 (2017) 181
Ramos et al. Entropy 20 (2018) 208
Hannig et al. FSI Gen 7 (2019) 572



Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Measuring calibration of LR-systems

- Metrics: summary statistics that measure calibration for the LR-system as a whole
  - Cllr_cal (Brummer and Du Preez, CSL 20 (2006) 230;
    Ramos et al, Entropy 20 (2018) 208)

  - Rates of misleading evidence: $P(LR \geq k|H_d) \leq \frac{1}{k}$ and $P\left(LR \leq \frac{1}{k}|H_p\right) \leq \frac{1}{k}$ for k = 2
    (Royall, "Statistical evidence: a likelihood paradigm")
    - Metrics: $\frac{\sum_{H_d} \mathrm{I}_{LR \geq 2}}{m}$ and $\frac{\sum_{H_p} \mathrm{I}_{LR \leq \frac{1}{2}}}{n}$ ("MislHd" and "MislHp")

  - Moments of LR-distributions: $\mathrm{E}\left(LR^n|H_p\right) = \mathrm{E}(LR^{n+1}|H_d)$ (Good, Bayes. Stat. 2 (1985) 249)
    - $1 = \mathrm{E}(LR|H_d)$ → Metric: $\frac{\sum_{H_d} LR}{m}$ ("Mom0")
    - $\mathrm{E}\left(\frac{1}{LR}|H_p\right) = 1$ → Metric: $\frac{\sum_{H_p} LR^{-1}}{n}$ ("Mommin1")

  - devPAV (Vergeer et al, FSI 321 (2021) 110722)

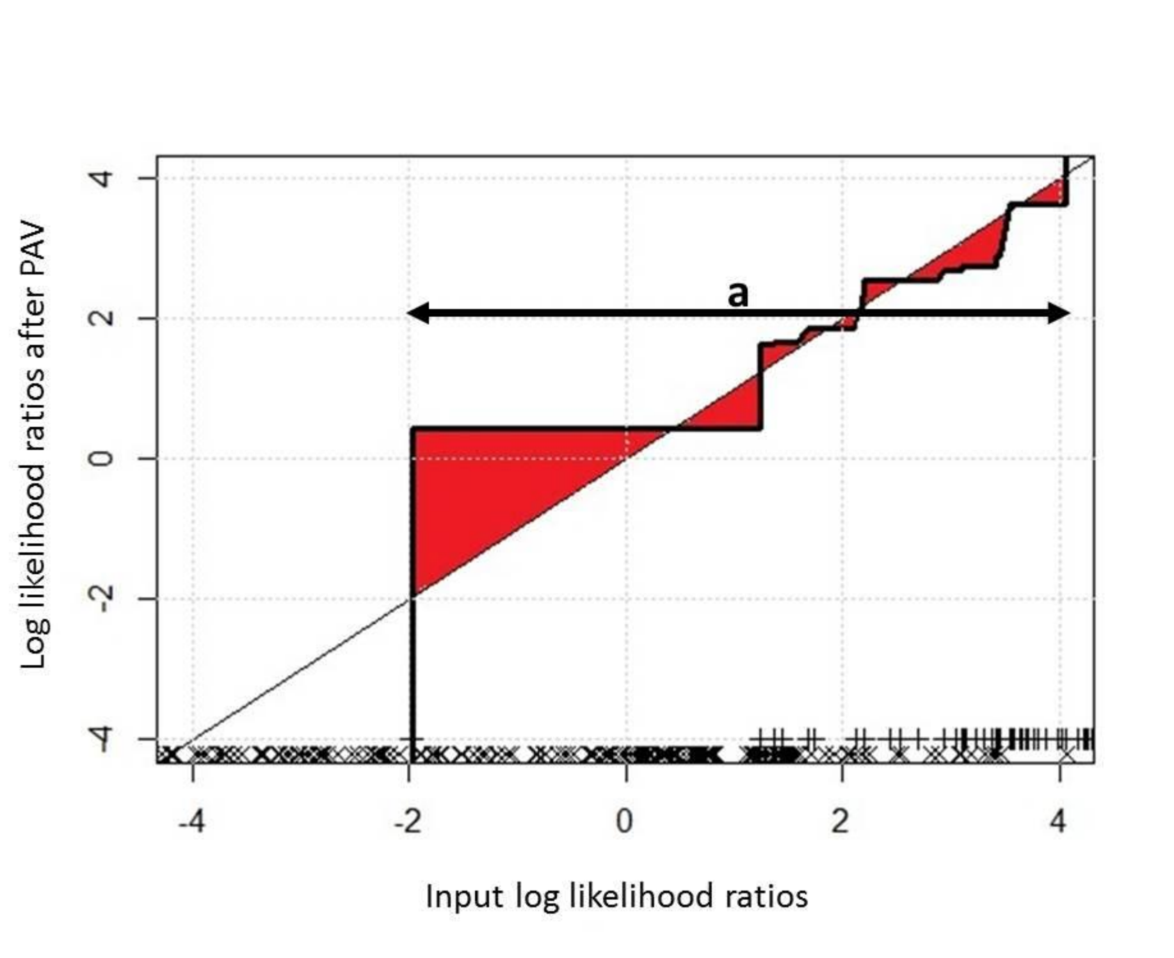
Measuring calibration of LR-systems
- devPAV

the red surface area
divided by the length of spline 'a'.

Remember from physics class:

$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{\int_t \; v(t)dt}{\Delta t}$$

devPAV = 'The average deviation of the PAV-transform from the identity line'



Random LLRs drawn from well- and ill-calibrated data

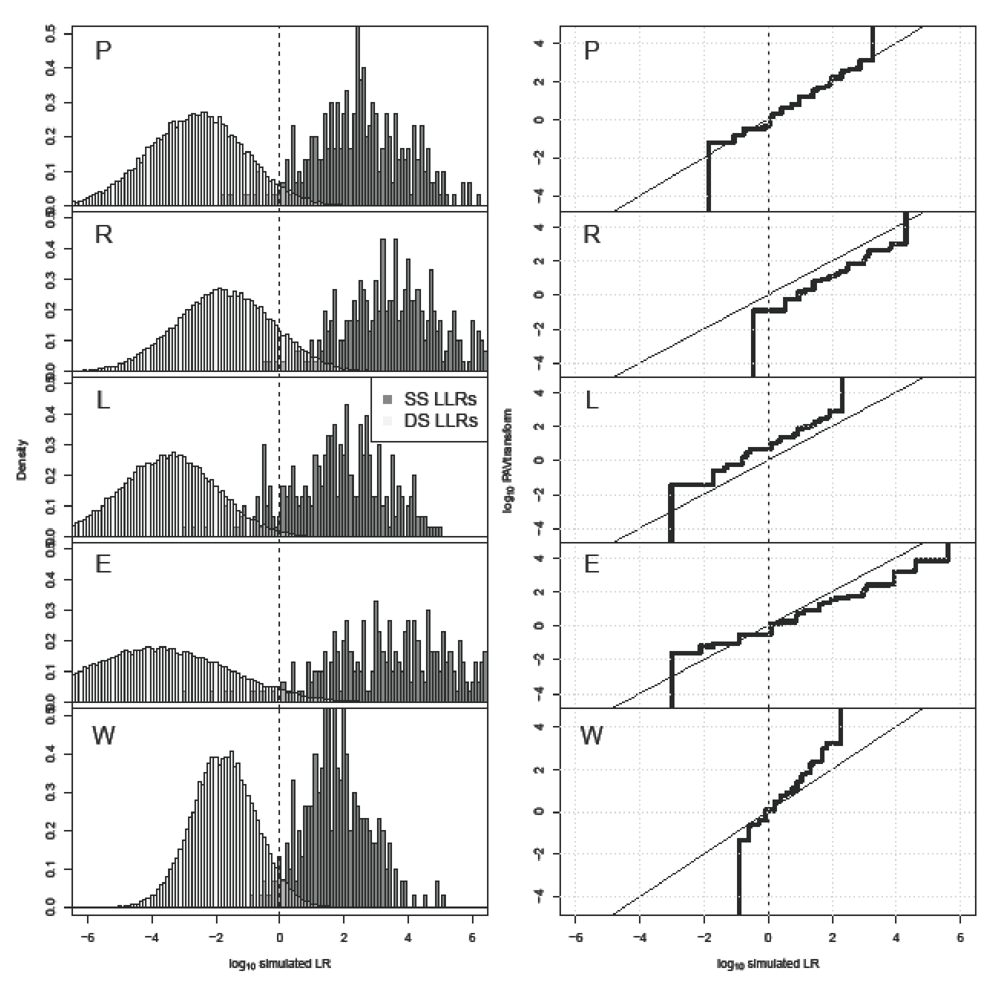
- Assumption: well-calibrated LLR-distributions are normal

$SS \sim \mathrm{N}(\mu_s, \sigma_s)$, and

$DS \sim \mathrm{N}(\mu_d, \sigma_d)$,

$\sigma := \sigma_s = \sigma_d$ and $\mu_s = -\mu_d$ applies.

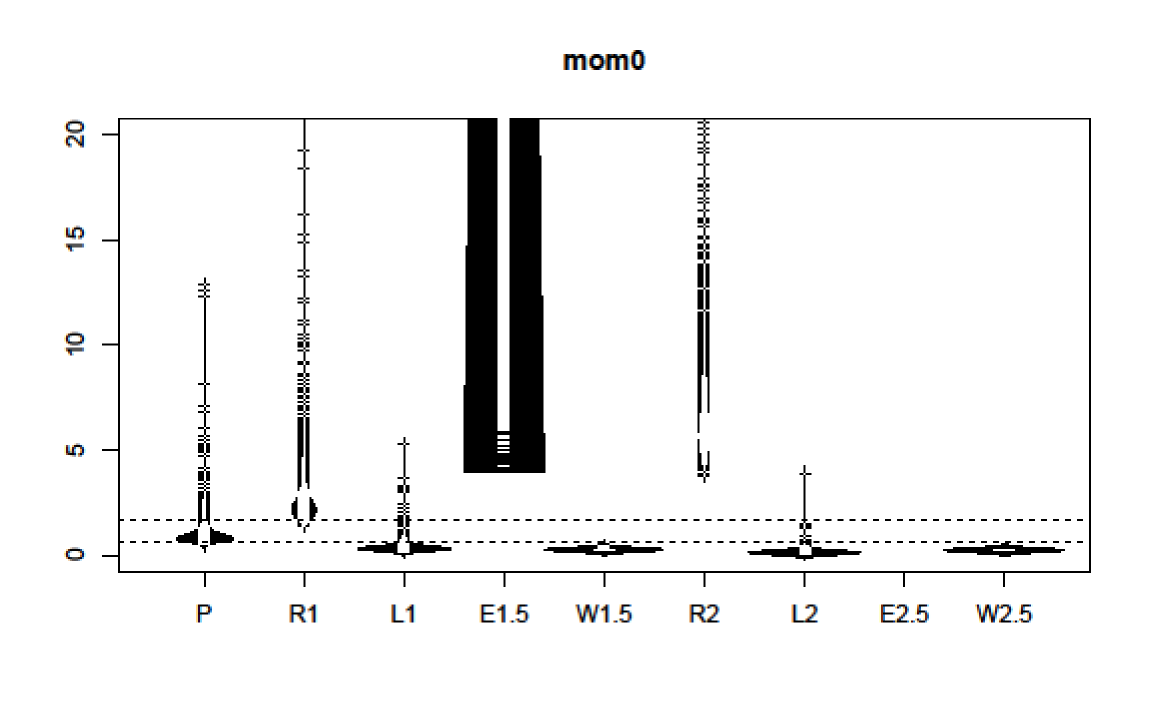
P → perfectly calibrated
R → to all LLRs a constant C (C > 0) is added
L → to all LLRs a constant C (C > 0) is subtracted
E → all LLRs are multiplied by C (C > 1)
W → all LLRs are divided by C (C > 1)

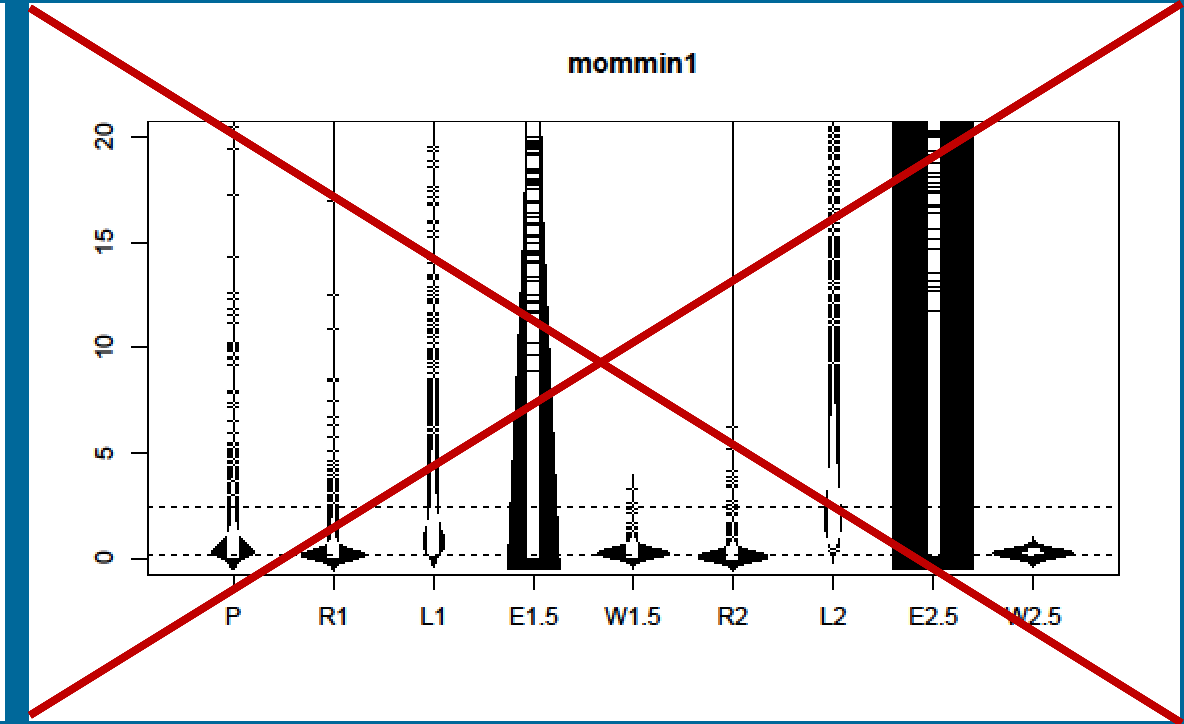


Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Mom0 and mommin1

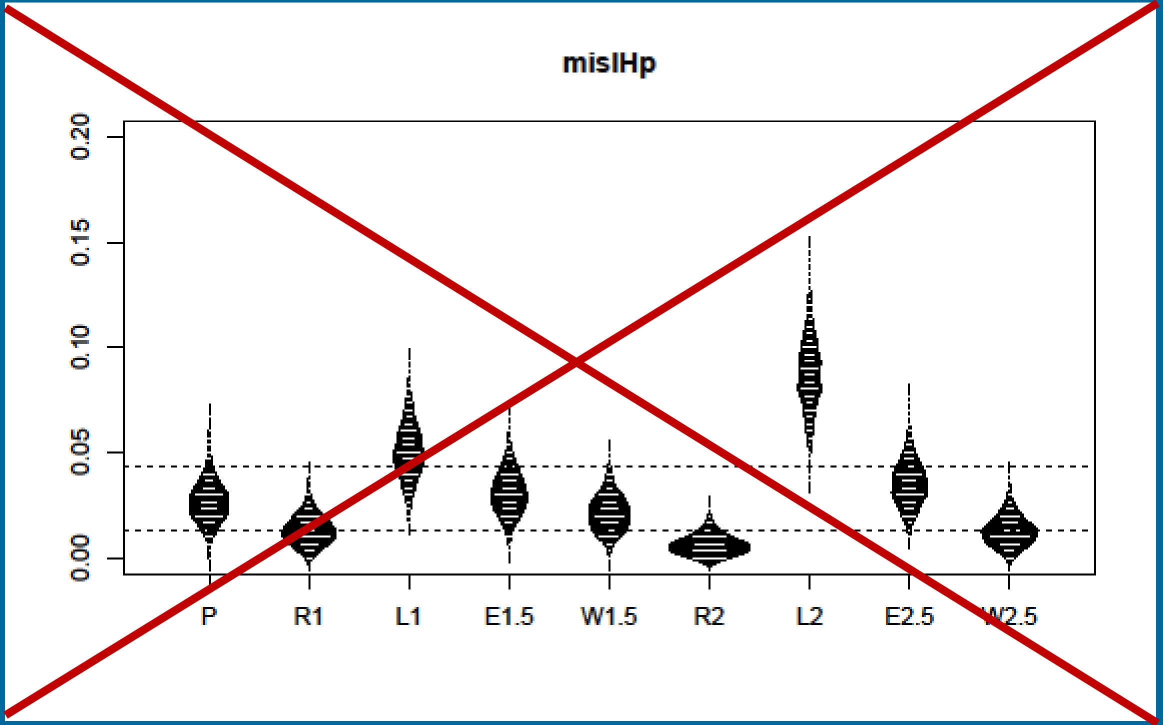
N_SS = 300, N_DS = $\frac{300 \times 299}{2}$, μ_s = 6 (EER = 4.2%)

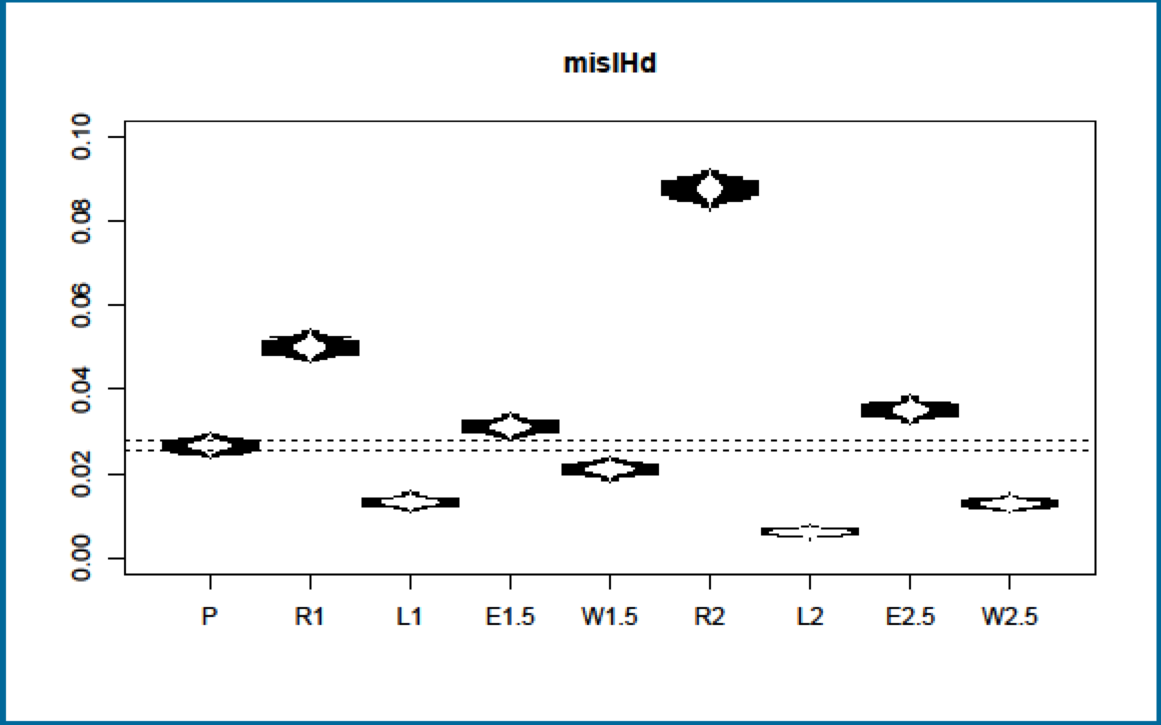




MislHp and MislHd

N_SS = 300, N_DS = $\frac{300 \times 299}{2}$, μ_s = 6 (EER = 4.2%)

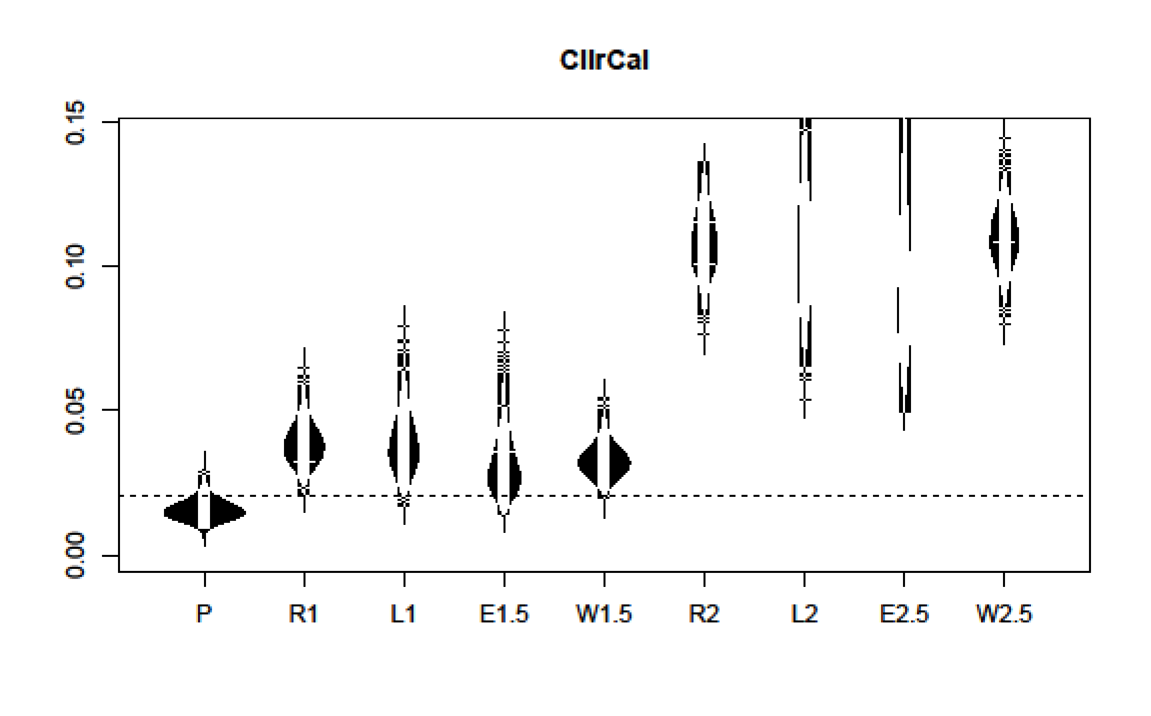


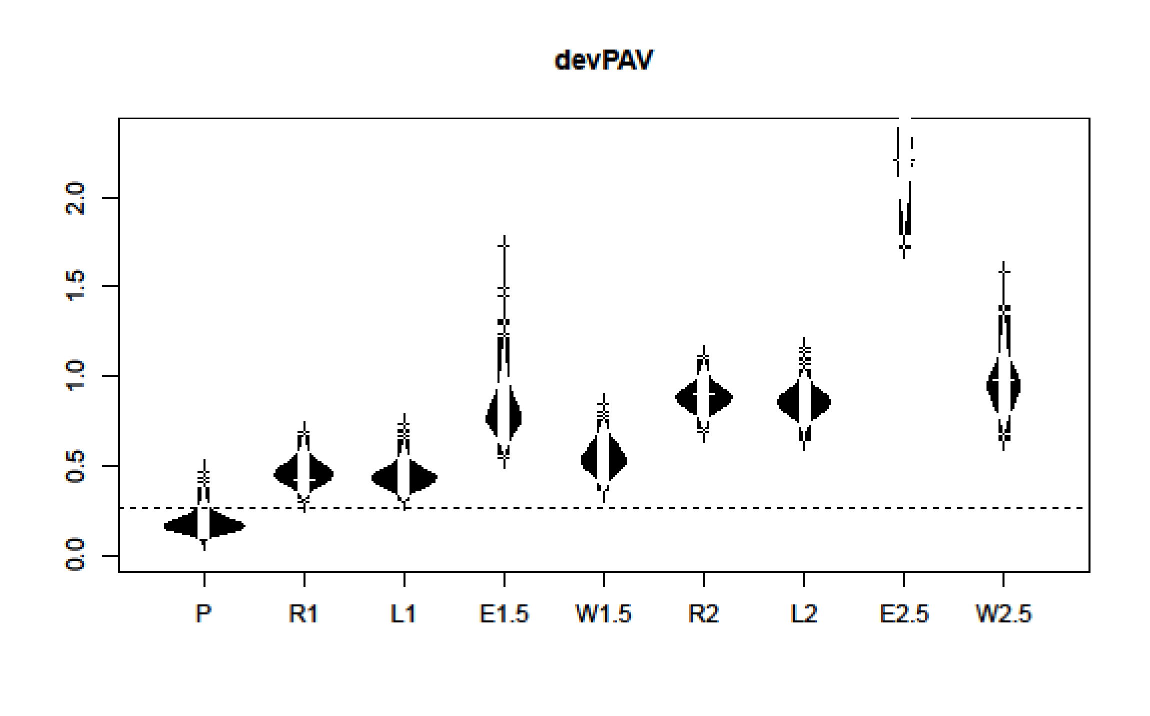


Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Cllr_cal and devPAV

N_SS = 300, N_DS = $\frac{300 \times 299}{2}$, μ_s = 6 (EER = 4.2%)

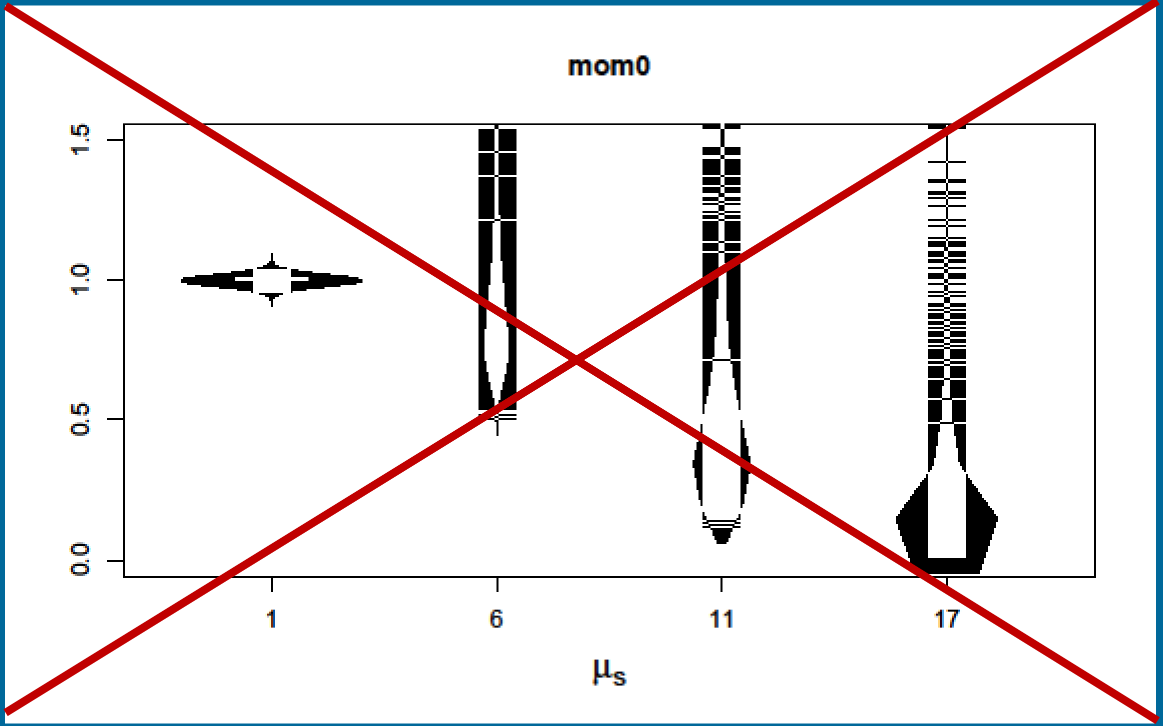


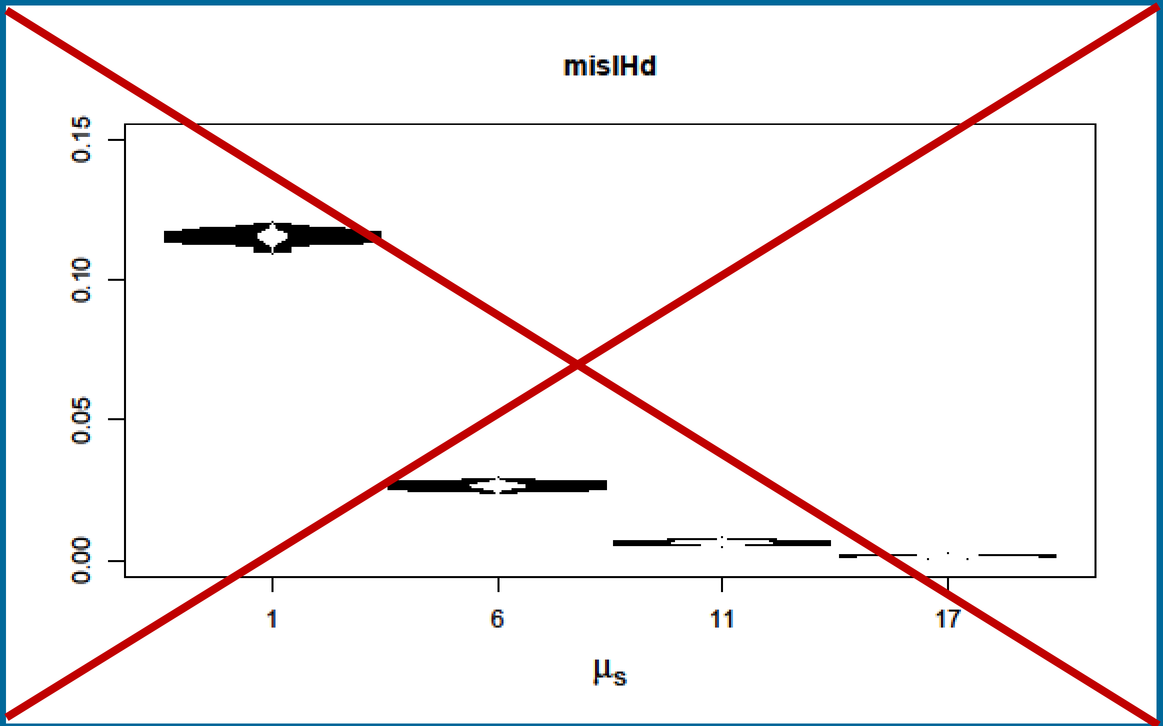


Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Stability: mom0 and mislHd

N_SS = 300, N_DS = $\frac{300 \times 299}{2}$, μ_s varied

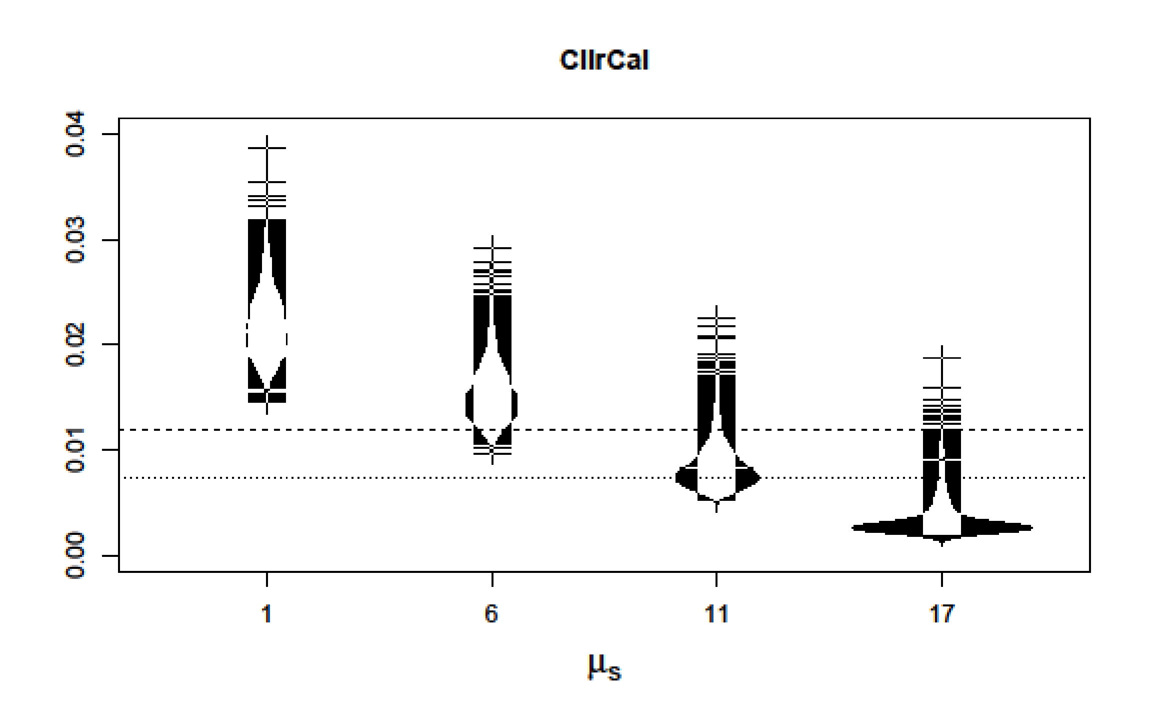


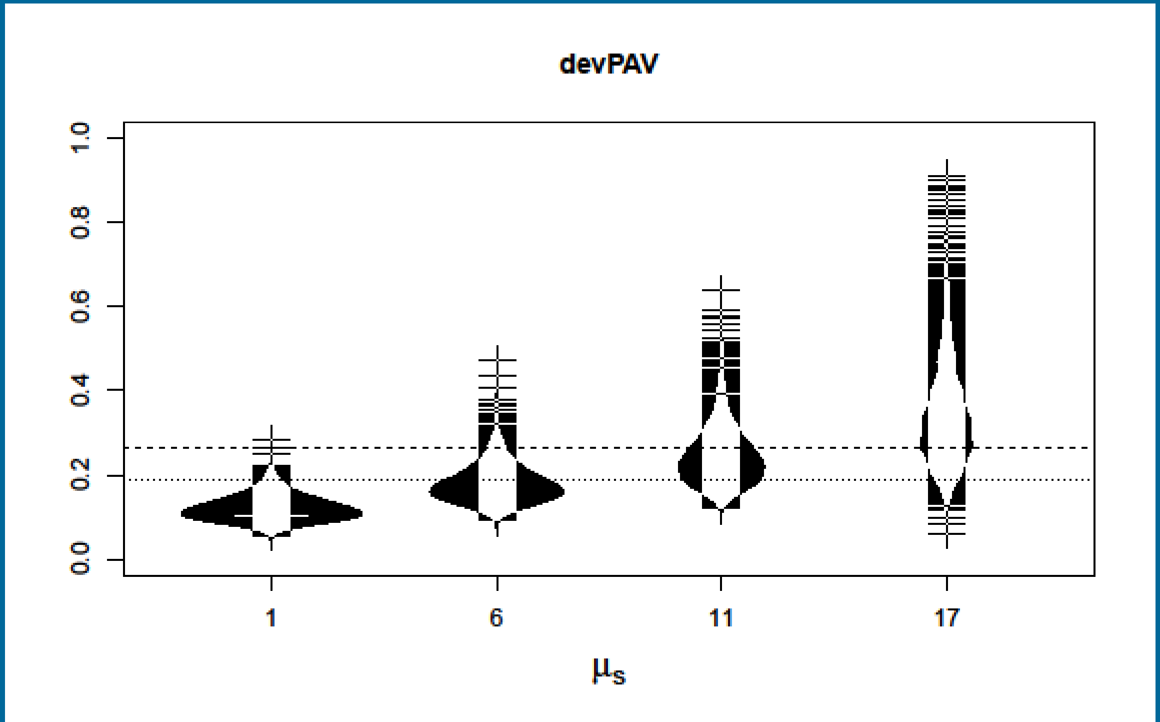


Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Stability: Cllr_cal and devPAV

N_SS = 300, N_DS = $\frac{300 \times 299}{2}$, $\mu_s$varied





Conclusion

- Measuring calibration is necessary to ensure that an LR-system adds information over the prior odds
- Visual representations of calibration are all based on binning and reading off relative frequencies
- Several metrics are used in the literature to measure calibration of LR-systems
  - Studied rates of misleading evidence and moments of LR-distributions lack differentiating abilities or are relatively unstable
  - Cllr_cal and devPAV differentiate well and are relatively stable
  - All metrics become unstable at some point when increasing discrimination of LR-systems