The empirical calibration of effect estimators in meta-analysis

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

Christian Röver and Tim Friede

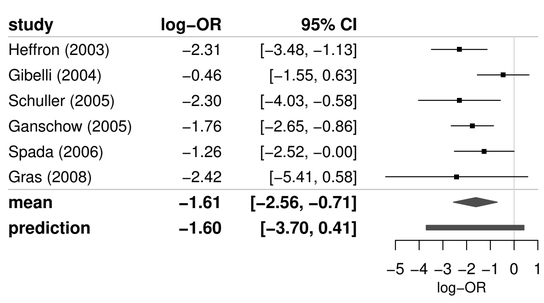
University Medical Center Göttingen, Germany

Meta-analysis

Individual **estimates** y_i (given along with **standard errors** s_i) are to be combined in a pooled analysis. The common **random-effects model** may be stated as:

$$y_i|\theta_i, s_i \sim \text{Normal}(\theta_i,\ s_i^2),$$
$$\theta_i|\mu, \tau \sim \text{Normal}(\mu,\ \tau^2)$$

for $i = 1, \ldots, k$. Interest usually is in **estimating** μ or in **predicting** θ_{k+1}.



The **heterogeneity** τ then constitutes a **nuisance parameter**.

Different approaches

Many analysis approaches are available within the random-effect model framework, [e.g., 1, 2, 3]:

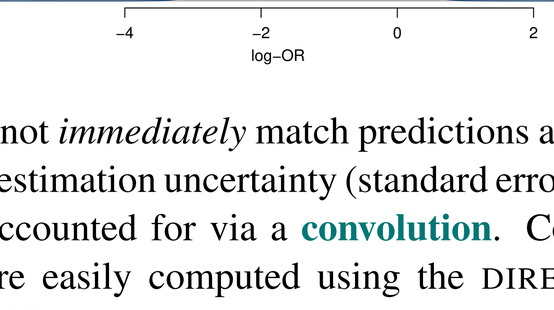
- **Bayesian** with different heterogeneity priors,
- **frequentist** with different types of confidence interval adjustment.

The Cochrane library

. . . contains data from **many** archived meta-analyses from medical applications. Question is whether we can utilize this rich data set to check and compare the proper **calibration** [4] of analysis methods.

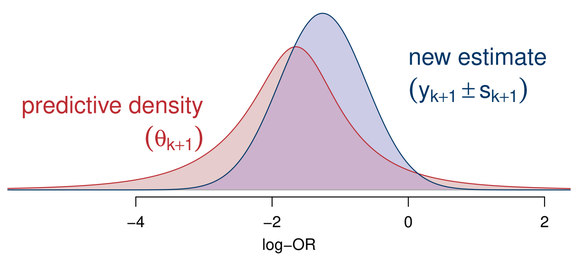
Leave-one-out calibration check

For a single analysis, we can try to **predict** one of the studies based on the remaining ones.



Repeated matching of **prediction** and actually **observed data** then should allow to evaluate whether the analysis is overall consistent.

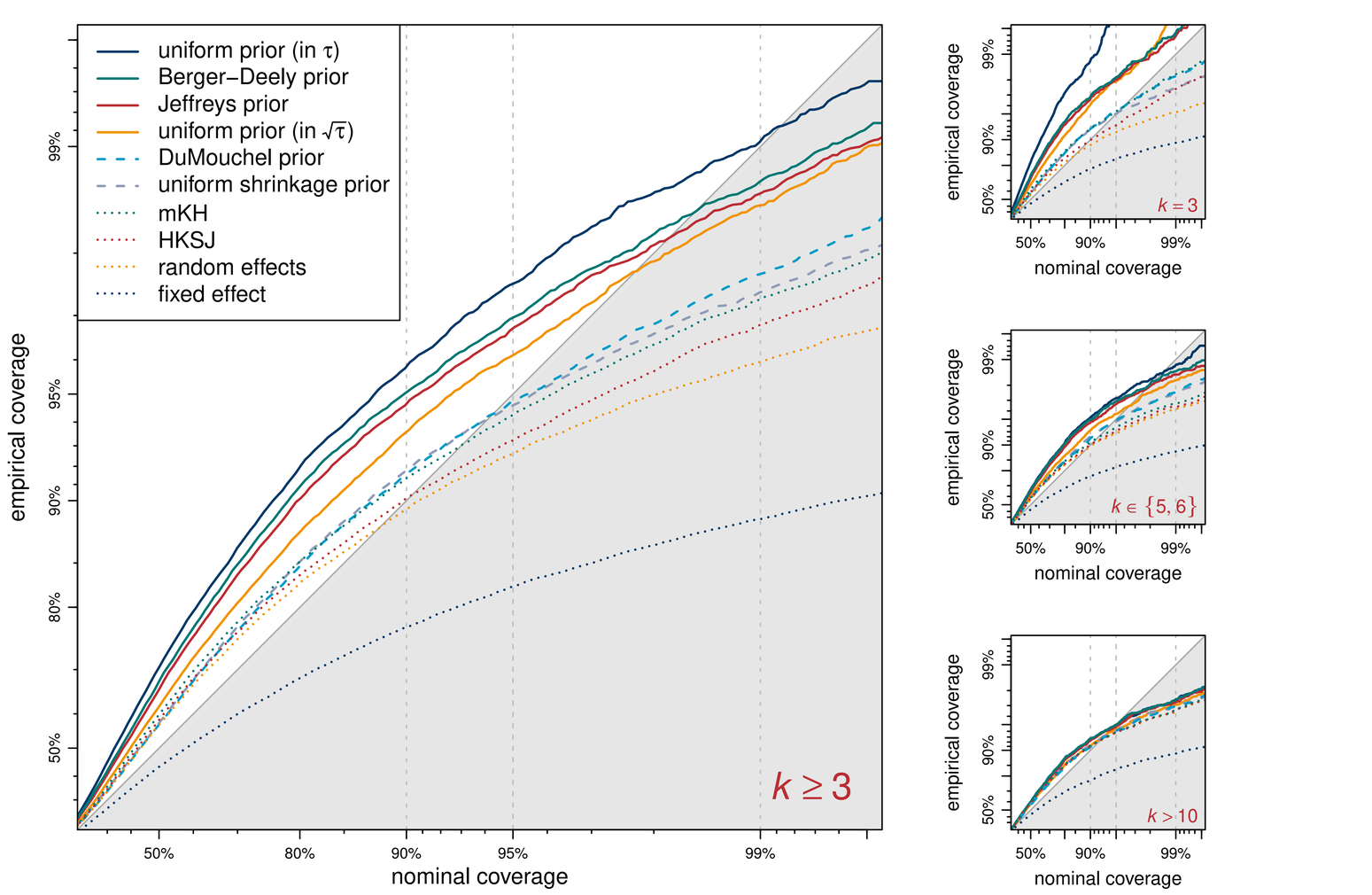
In the leave-one-out step, we do not know the **true values** (θ_{k+1}), but only an **uncertain estimate** (y_{k+1} with standard error s_{k+1}).



We cannot *immediately* match predictions and estimates; estimation uncertainty (standard error) needs to be accounted for via a **convolution**. Convolutions are easily computed using the DIRECT approach [5].

Matching predictions and observables

By adding estimation uncertainty on top of the predictive distribution, we can effectively derive prediction intervals for **observations** y_{k+1} (rather than the **true values** θ_{i+1}). With that, we may then empirically investigate the calibration (coverage) of different inference methods [4]. Here we focus on Bayesian estimates using an (improper) uniform effect and **noninformative heterogeneity priors**, and some **common frequentist approaches**.



The calibration of meta-analysis methods across the Cochrane Library data (overall and for different meta-analysis sizes k).

Calibration

The investigation includes 19 500 published meta-analyses and yields the resulting calibrations for varying nominal credible / confidence levels (see figure). Several improper (*uniform*, *Berger-Deely*, *Jeffreys*) and proper (*DuMouchel*, *uniform shrinkage*) non-informative priors are considered [2]. The investigated frequentist models include random-effects models utilizing a normal approximation as well as *Hartung-Knapp-Sidik-Jonkman (HKSJ)* and *modified Knapp-Hartung (mKH)* intervals [3].

All methods fail to reach nominal coverage at high nominal levels. Among the **Bayesian methods**, the improper priors yield at least conservative coverage at the 95% level. The **frequentist methods** perform best when using HKSJ or mKH adjustment, but still do not reach the 95% level. All methods (except for the fixed-effect model) tend to perform similarly for large k.

Beyond calibration

Calibration is a necessary condition for probabilistic inference methods. Beyond that, different *calibrated* methods may still differ in their precision (**sharpness**) [4], which we are currently investigating based on **scores**.

Conclusions

Bayesian methods based on uninformative priors work well for many studies (large k), and tend to be conservative for few studies. In the latter case, the use of weakly informative heterogeneity priors may be more appropriate [6].

The miscalibration at high nominal levels suggests that a heavier-tailed heterogeneity model component may be more appropriate (this is also supported by additional simulations, not shown here).

References

[1] W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), August 2010. doi: 10.18637/jss.v036.i03.

[2] C. Röver. Bayesian random-effects meta-analysis using the bayesmeta R package. *arXiv preprint 1711.08683*, November 2017. URL http://www.arxiv.org/abs/1711.08683.

[3] C. Röver, G. Knapp, and T. Friede. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*, 15, November 2015. doi: 10.1186/s12874-015-0091-1.

[4] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society B*, 69(2):243–268, April 2007. doi: 10.1111/j.1467-9868.2007.00587.x.

[5] C. Röver and T. Friede. Discrete approximation of a mixture distribution via restricted divergence. *Journal of Computational and Graphical Statistics*, 26(1):217–222, February 2017. doi: 10.1080/10618600.2016.1276840.

[6] T. Friede, C. Röver, S. Wandel, and B. Neuenschwander. Meta-analysis of few small studies in orphan diseases. *Research Synthesis Methods*, 8(1):79–91, March 2017. doi: 10.1002/jrsm.1217.