Bayesian model building: Hierarchical modeling and meta analysis

Katja Ickstadt, TU Dortmund University

joint work with

Jochem König (IMBEI Mainz) and Gerhard Nehmiz (Consultant, Boehringer Ingelheim)

Symposium: Recent Advances in Meta-Analysis

Göttingen 2024

technische universität
dortmund

RTG 2624

Overview

1. Comparison of Bayesian and frequentist statistics, Part 1
2. Hierarchical modeling: An example
3. Hierarchical models
4. Estimation and prediction
5. Comparison of Bayesian and frequentist statistics, Part 2

Comparison of Bayesian and frequentist statistics, Part 1

Comparison of Bayesian and frequentist statistics

Definitions of probability

- **Frequentist statistics:**

Probability as the relative frequency with which an event occurs in a large number of identical, repeated, independent random experiments.

- **Bayesian statistics:**

Probability as the degree of certainty for a statement (e.g. a range for a relative proportion, a range for an intervention effect), given observations/data

Comparison of Bayesian and frequentist statistics

Remarks:

- **Interpretation:**
  Posterior distributions and derived quantities are easier to interpret than their frequentist counterparts (e.g., credibility intervals versus confidence intervals, tail probabilities versus p-values).
- **For a long time:**
  Dominance of frequentist statistics because less computationally intensive

Hierarchical modeling: An example

Hierarchical models in general:

Multiparameter models where parameters are structurally dependent. The joint distribution of all parameters reflects these dependencies.

Example: PRoMPT-study

PRoMPT-study (“PRimary care Monitoring for depressive Patients Trial“, Gensichen et al., 2005 and 2009):

Background

Case management by health care assistants in small primary care practices provides unclear benefit for improving depression symptoms.

Objective

To determine whether case management provided by health care assistants in small primary care practices is more effective than usual care in improving depression symptoms and process of care for patients with major depression.

Example: PRoMPT-study

Design

Cluster randomized, controlled trial. The practices form the clusters.

Same and similar models apply in meta analyses.

Setting

74 small primary care practices in Germany from April 2005 to September 2007; 39 control practices (standard); 35 case management practices with 1 practice that did not recruit any patients.

Example: PRoMPT-study

Patients

626 patents (316 in control practices, 310 in case management practices), age 18 to 80 years, with major depression.

Intervention

Structured telephone interviews by healthcare assistants to monitor depression symptoms and support for adherence to medication, with feedback to the family physician.

Measurements

Outcome: Depression symptoms at 12 months, as measured by the Patient Health Questionnaire-9 (PHQ-9); *PHQ-9 score:* Sum of 9 variables, each given on a Likert scale with 4 possible answers (0= never,..., 3= almost daily); assumed to be normally distributed

Example: PRoMPT-study

Data base

- 272 patients of the 39 control practices and 242 patients of the 34 case management practices with baseline and a 12 months PHQ-9 scores.
- 44 and 68 patients, respectively, showed a baseline score only
- 16 and 25 patients, respectively, showed a baseline and an additional 6-months score

→ This 6-months score with an additional correlation assumption yields an estimation for an unobserved 12-months score, leading to the data base of 288 patients in 39 control practices and 267 patients in 34 case management practices for the primary analysis in Gensichen et al. (2009).

→ **now analyzed:** 272 patients of the 39 control practices and 242 patients of the 34 case management practices with baseline and a 12 months PHQ-9 scores.

Example: PRoMPT-study

Assumptions

- Patients within each practice are similar (with respect to, e.g., age, sex, social status)
- Patients between practices could possibly show differences in prognostic relevant features, treatment quality/treatment specificities, (and, perhaps, in the outcome variable)
- Y_{ij} random variables for measurement of patient j in practice i
- Outcome expectations: A_i

→ Practice affiliation might explain part of outcome variable's variance

→ Practices are viewed as clusters of patients and as a random draw of a population of practices

→ Practices are assumed independent and randomized in 2 groups (control (=0) and case management (=1))

→ Hierarchy: Distribution of the A_i determines distribution of the Y_{ij} but not vice versa.

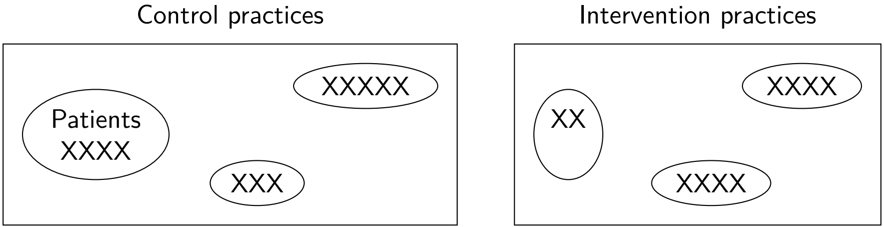

Figure 1: Hierarchical structure

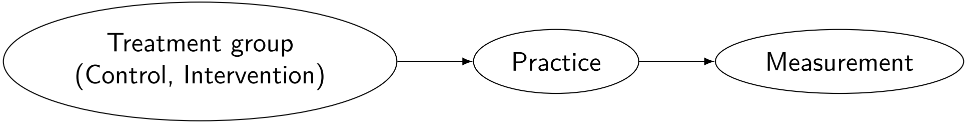

Figure 2: Influence diagram

Hierarchical models

Important hierarchical model class in general

Linear Mixed Model, LMM:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

The design matrices $\mathbf{X}(n \times p_1)$ and $\mathbf{Z}(n \times p_0)$ are given with p_1 fixed effects and p_0 random effects parameters. Further:

$$\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{R})$$

and

$$\boldsymbol{\gamma} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{G})$$

with given covariance matrices $\mathbf{G}(p_0 \times p_0)$ for random effects, $\boldsymbol{\gamma}(p_0 \times 1)$ and $\mathbf{R}(n \times n)$ for the residual errors $\boldsymbol{\varepsilon}(n \times 1)$ which are independent from them.

The **linear predictor** is defined as $\mathbf{X}\beta + \mathbf{Z}\boldsymbol{\gamma}$.

Important hierarchical model class in general

Assuming normal distributions for γ and ε, we can write the model hierarchically as:

$$\mathbf{Y}|\boldsymbol{\gamma} \sim \mathrm{MVN}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \mathbf{R}),$$

$$\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} \sim \mathrm{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZGZ}^\top).$$

Conditional on each fixed $\boldsymbol{\beta}$ we obtain marginally:

$$\mathbf{Y} \sim \mathrm{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{R} + \mathbf{ZGZ}^\top),$$

the marginal variance-covariance matrix $\mathbf{R} + \mathbf{ZGZ}^\top$ is also called $\mathbf{V}$. The expected value of $\mathbf{Y}$ is $\mathbf{X}\boldsymbol{\beta}$.

Hierarchical model for PRoMPT example

For the PRoMPT-study example:

- $\boldsymbol{\beta} = \binom{\mu_0}{\mu_1}$ mean values of $p_1 = 2$ treatment groups,
- $\mathbf{X} = \left(\begin{array}{c|c} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{array}\right)$ the indicators for $p_1 = 2$ treatments,

Hierarchical model for PRoMPT example

- $\mathbf{Z} = \left(\begin{array}{c|ccc} 1 & & & \\ \vdots & & 0 & \\ 1 & & & \\ \hline & 1 & & \\ & \vdots & & \\ & 1 & & \\ 0 & & \ddots & \\ & & & 1 \\ & & & \vdots \\ & & & 1 \end{array}\right)$
  a matrix with blocks $J_{n_i \times 1}$ for the assignment of patients to medical practices, and
- $\boldsymbol{\gamma} = (a_1, \ldots, a_k)$ the indicators for the expected values in the p_0 medical practices (both treatments together), centered on 0. Due to the independence between $\boldsymbol{\gamma}$ and all ε_i
- $\mathbf{R}$ and $\mathbf{G}$ also have a block structure

Hierarchical model for PRoMPT example

I.e.:

- $\mathbf{R} = \sigma^2 I_n$,
- $\mathbf{G} = \tau^2 I_{p_0}$ and
- $$\mathbf{V} = \mathbf{R} + \mathbf{ZGZ}^T = \left(\begin{array}{cccc|c} \tau^2+\sigma^2 & \tau^2 & \cdots & \tau^2 & \\ \tau^2 & \ddots & & & 0 \\ \vdots & & \ddots & & \\ \tau^2 & & & \tau^2+\sigma^2 & \\ \hline & & 0 & & \cdots \end{array}\right)$$

  the marginal variance-covariance matrix with 1 block per medical practice.

Hierarchical model for PRoMPT example

We assume a normal distribution for the expected value of practice (cluster) i in each treatment group if the practices are independent given their treatment group:

$$A_i \sim \mathrm{N}(\mu_0, \tau^2), i = 1, \ldots, k_0 \quad \text{i.i.d.}$$

resp.

$$A_i \sim \mathrm{N}(\mu_1, \tau^2), i = k_0 + 1, \ldots, k_0 + k_1 \quad \text{i.i.d.},$$

where the variance τ^2 of A_i is equal for all practices and, in particular, is independent of the treatment group.

The distribution for A_i for both treatment groups are called **population distributions**.
Parameters of these population distributions are called **hyperparameters**, in our example: τ, μ_0 and μ_1.
Population distributions **combine information** across units (pooling).

Hierarchical model for PRoMPT example

At the patient level, the individual values are assumed to be identically normally distributed, independently around the practice expectation value $A_i = a_i$, with standard deviation σ:

$$Y_{ij} = a_i + \varepsilon_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, k_0 + k_1,$$

where all individual deviations $\varepsilon_{ij} = Y_{ij} - a_i$ are independent of each other and of all practice expectations A_i.

The actual expectation values of the practices a_i are not observable (**latent**). Only the measured values y_{ij} themselves can be observed.

Hierarchical model for PRoMPT example

In summary the hierarchical model can be specify as follows:

$$Y_{ij}|A_i = a_i \sim \mathrm{N}(a_i, \sigma^2)$$

and

$$A_i \sim \mathrm{N}(\mu_0, \tau^2) \text{ for control practices } (i = 1, \ldots, k_0) \text{ resp.}$$

$$A_i \sim \mathrm{N}(\mu_1, \tau^2) \text{ for intervention practices } (i = k_0 + 1, \ldots, k_0 + k_1)$$

The parameters in this model are: $\sigma^2, \mu_0, \mu_1, \tau^2$ and a_i, where the distributions of A_i are conditional on μ_0, μ_1 and τ.

Hierarchical model for PRoMPT example

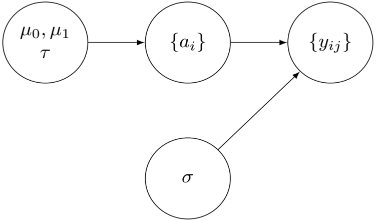

Figure 3: Directed acyclic graph for data and parameters

The A_i are independent given the hyperparameters. With the product representation for the prior distribution of all $A_i, \sigma, \mu_0, \mu_1, \tau$, the joint prior distribution is invariant with respect to the permutation of the A_i. This invariance carries through to the posterior distribution, and is called **exchangeability**.

Estimation and prediction

Classical analysis for PRoMPT example

Only a_i and σ^2 are to be estimated (cf ANOVA). → ML-estimates:

$$\hat{a_i}^{(ML)} = \bar{y}_i \text{ and } \hat{\sigma^2}^{(ML)} = 1/n \sum_{i,j} (y_{ij} - \bar{y}_i)^2 .$$

I.e., if we consider the a_i as fixed parameters (more generally: the conditional likelihood, given the random effects, for an ML estimation) then the ML estimator for a_i is equal to $\bar{y}_i$.

Further, if μ_0 is fixed but unknown and σ and τ are known: Then, according to DerSimonian and Laird (1986), the ML-estimate is the weighted mean:

$$\hat{\mu_0}^{(\mathrm{ML})} = \frac{\sum_i \frac{1}{\frac{\sigma^2}{n_i}+\tau^2} \bar{y_i}}{\sum_i \frac{1}{\frac{\sigma^2}{n_i}+\tau^2}}.$$

Classical analysis for PRoMPT example

The conditional expected value $\mathbb{E}(A_i|\mathbf{Y}=\mathbf{y})$:

For each given $\mu_0, \mu_1, \sigma, \tau$ we obtain

$$\mathbb{E}(A_i|\mathbf{Y}=\mathbf{y}) = \mathbb{E}(A_i|\bar{y_i}) = \mu_0 + (\bar{y}_i - \mu_0)w_i$$

with weights:

$$w_i = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n_i}} \in [0,1]\,.$$

Thus the $\hat{A}_i$ are shrunk towards the joint mean μ_0. The same applies to μ_1.

Classical analysis for PRoMPT example

Properties for w_i:

- w_i is practice-specific
- If the intraclass correlation coefficient (ICC) is defined as $\frac{\tau^2}{\tau^2+\sigma^2}$, then $w_i = \frac{n_i \cdot ICC}{1+(n_i-1)\cdot ICC}$. $\rightarrow \mathbb{E}(A_i|\bar{y_i}) - \mu_0$ for practice i is therefore reduced compared to the "raw" difference $(\bar{y}_i - \mu_0) \rightarrow$ **'Shrinkage'**
- w_i close to 1, if τ is large in relation to σ i.e. the practices are heterogeneous, or if the number of patients n_i of the practice i is large.
- w_i close to 0 if both the practices are homogeneous (τ small) and there are only a few observations in the practice.

Bayesian estimation and prediction for PRoMPT example

Bayesian inference for analyzing the hierarchical model:

Specification of prior distributions:

- weakly informative prior distributions for a_i and σ^2:

$$A_i \sim \mathbf{N}(0, 10^6) \quad (i = 1, \ldots, k_0 + k_1), \quad \frac{1}{\sigma^2} \sim \texttt{Gamma}(0.001, 001)$$

- prior distributions for the hyperparameters:

$$\mu_0 \sim \mathbf{N}(0, 10^6), \quad \mu_1 \sim \mathbf{N}(0, 10^6), \quad \tau \sim \texttt{unif}(0, M) \text{ for large } M$$

Remark: Alternative for the scale parameter τ^2

$$\tau^2 \sim \text{half-t with } \geq 1 \text{ degrees of freedom.}$$

Bayesian analysis

- Gibbs sampling for Linear Mixed Models (e.g., Gelfand and Sahu, 1999)
- 2 chains for each parameter, with 100000 iterations following 4000 iterations of burn-in
- Monte Carlo error less than 5% of posterior standard deviation

	Median (95% CrI)
intervention	$10.75 \quad (9.87, 11.64)$
control	$12.14 \quad (11.30, 12.97)$
intervention - control	$-1.39 \quad (-2.59, -0.16)$
σ	$5.66 \quad (5.31, 6.05)$
σ^2	$32.1 \quad (28.2, 36.6)$
τ	$1.42 \quad (0.55, 2.18)$
τ^2	$2.00 \quad (0.30, 4.75)$
ICC	$0.059 \quad (0.009, 0.133)$

Table 1: MCMC estimation of μ_1, μ_0, the difference $\mu_1 - \mu_0$,variance components and the ICC. Intervention: $n = 242$ in 34 practices. Control: $n = 272$ in 39 practices

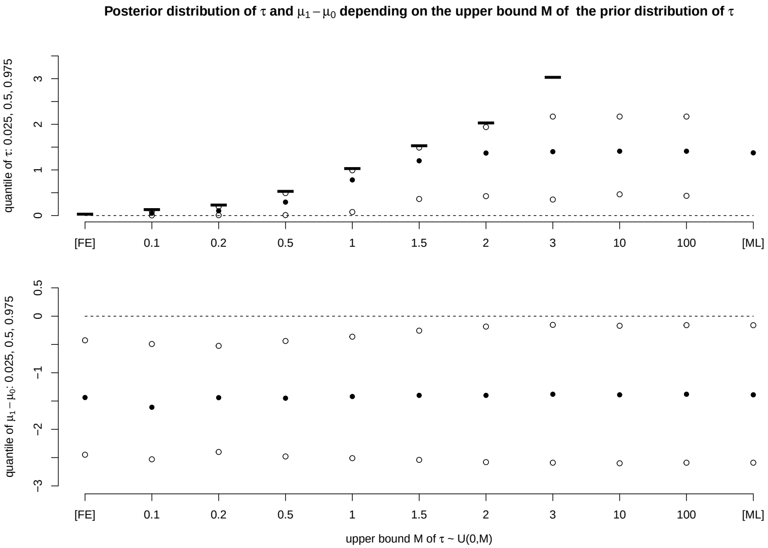

Figure 4: Sensitivity analysis

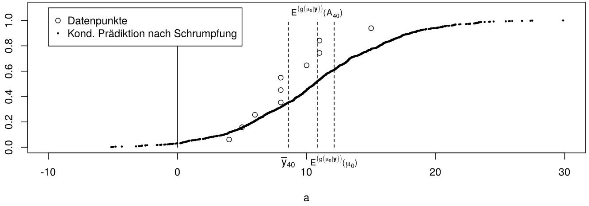

Figure 5: Conditional prediction of the measurement of a new patient in practice 40 (control), probability for $\tilde{Y}_{40} \leq a$

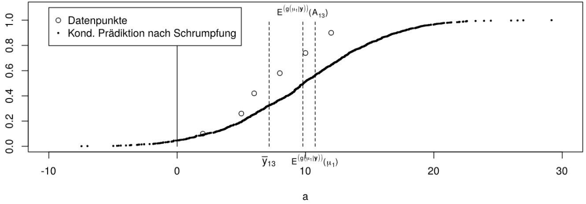

Figure 6: Conditional prediction of the measurement of a new patient in practice 13 (intervention), probability for $\tilde{Y}_{13} \leq a$

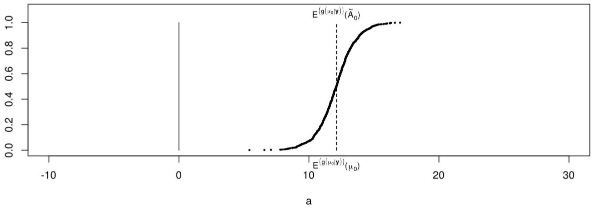

Figure 7: Prediction of the expected value of a new control practice, probability for $\tilde{A}_0 \leq a$

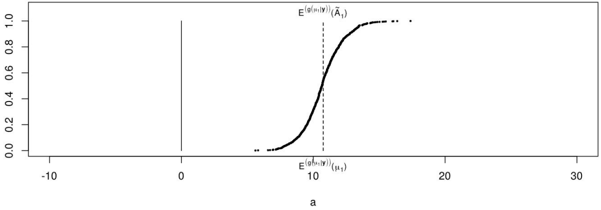

Figure 8: Prediction of the expected value of a new intervention practice, probability for $\tilde{A}_1 \leq a$

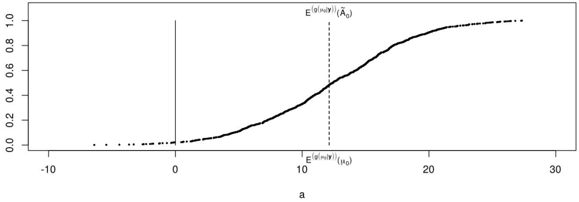

Figure 9: Prediction of the measurement of a new patient in new control practice, probability for $\tilde{Y}|\tilde{A}_0 \leq a$

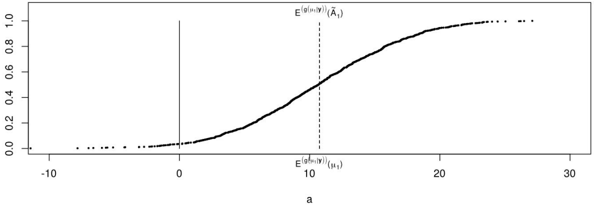

Figure 10: Prediction of the measurement of a new patient in new intervention practice, probability for $\tilde{Y}|\tilde{A}_1 \leq a$

Discussion

1. Meta analyses can be viewed as specific cases of analyzing hierarchical models
2. Here, exemplified for a cluster randomized trial
3. A weakness for a frequentist analysis of Linear Mixed Models is that usually τ is assumed to be known. In Bayesian inference we can easily treat τ as a parameter.
4. A sensitivity analysis for the formulation of the prior distribution of τ should be performed. In our example: the analysis is robust to the specific choice of the prior.
5. Prediction of new patients for a specific practice, but also for a new practice are straight forward within the Bayesian framework.
6. The whole framework can be extended to Generalized Linear Mixed Models.

Comparison of Bayesian and frequentist statistics, Part 2

Comparison of Bayesian and frequentist statistics

Including prior knowledge generally desirable: Even if informative prior knowledge is not desired/existent, weakly informative prior distributions have advantages:

- downweight unimportant parameter values
- remedy against overfitting
- elegant formulation of parameter constraints.

Accuracy of the posterior distribution:
Posterior distribution can be specified exactly except approximation error (e.g. Monte Carlo error). Hence, in Bayesian statistics large-sample approximations e.g. via the asymptotic normal distribution at the Maximum Likelihood estimators play a minor role compared to frequentist statistics.

Comparison of Bayesian and frequentist statistics

Nuisance Parameters:
Can be integrated out of the posterior distribution, which is hard in a frequentist setting. Again, remedy against overfitting.

Uncertainty Propagation:
Sampling procedures, e.g., MCMC approaches lead to natural uncertainty propagation for derived quantities.

Comparison of Bayesian and frequentist statistics

Avoiding null-hypothesis significance tests:
Posterior distributions allow connections to cost-benefit considerations and to rational decisions. Frequentist analyses often result in typical null-hypothesis significance testing, which is being criticized to an increasing degree (e.g., McShane et al., 2019). It is especially problematic for null-hypotheses consisting of only one point in parameter space. In Bayesian analyses null-hypothesis testing this is not as common as in frequentist analyses.

Posterior predictive checks (PPCs):
Part of the so-called Bayesian Workflow (Gelman et al., 2020); easy to perform and intuitive way of a Bayesian model diagnosis. Expertise needed, but model diagnosis is often difficult in a frequentist setting.

References

- Brown, H. and Prescott, R. (1999) Applied Mixed Models in Medicine. John Wiley & Sons, Ltd. 34-39
- DerSimonian, R., Laird, N. (1986). Meta-analysis in clinical trials. Controlled clinical trials, 7(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2
- Gelfand, A. E., Sahu, S. K. (1999). Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models. Journal of the American Statistical Association, 94(445), 247–253. https://doi.org/10.1080/01621459.1999.10473840
- Gelman, A., Vehtari, A., Simpsons, D., Margossian, C.C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.C. (2020). Bayesian Workflow. arxiv.org/abs/2011.01808.
- Gensichen, J., Torge, M., Peitz, M. et al. (2005). Case management for the treatment of patients with major depression in general practices – rationale, design and conduct of a cluster randomized controlled trial – PRoMPT (Primary care Monitoring for depressive Patient's Trial) [ISRCTN66386086] – Study protocol. BMC Public Health 5, 101. https://doi.org/10.1186/1471-2458-5-101
- Gensichen, J., von Korff, M., Peitz, M., Muth, C., Beyer, M., Güthlin, C., Torge, M., Petersen, J. J., Rosemann, T., König, J., Gerlach, F. M., & PRoMPT (PRimary care Monitoring for depressive Patients Trial) (2009). Case management for depression by health care assistants in small primary care practices: a cluster randomized trial. Annals of internal medicine, 151(6), 369–378. https://doi.org/10.7326/0003-4819-151-6-200909150-00001
- McShane, B. B., Gal, D., Gelman, A., Robert, C., Tackett, J. L. (2019). Abandon Statistical Significance. The American Statistician, 73(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253
- Neuenschwander, B., Capkun-Niggli, G., Branson, M., Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. Clinical trials (London, England), 7(1), 5–18. https://doi.org/10.1177/1740774509356002

Thank you for your attention!

Contact: ickstadt@statistik.tu-dortmund.de

Supported by the Research Training Group ”RTG2624:Biostatistical Methods for High-Dimensional Data in Toxicology” funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project Number 427806116)