

Methods for correlated and overdispersed count data for meta-analysis of binary outcomes



Oliver Kuß (joint work with Ferdinand Stoye and Annika Hoyer)

Deutsches Diabetes-Zentrum (DDZ), Leibniz-Zentrum für Diabetes-Forschung an der Heinrich-Heine-Universität Düsseldorf, Institut für Biometrie und Epidemiologie

Two new methods (and a provocative thesis) for the meta-analysis of binary outcomes



Oliver Kuß (joint work with Ferdinand Stoye and Annika Hoyer)

Deutsches Diabetes-Zentrum (DDZ), Leibniz-Zentrum für Diabetes-Forschung an der Heinrich-Heine-Universität Düsseldorf, Institut für Biometrie und Epidemiologie

The mother of all meta-analyses in diabetes research



German Diabetes Center

P People with Type 2 Diabetes
I Rosiglitazone
C No rosiglitazone
O Number of cardiovascular deaths
S 42 RCTs



Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes

Steven E. Nissen, M.D., and Kathy Wolski, M.P.H.

- “In the rosiglitazone group, as compared with the control group, the odds ratio [...] for death from cardiovascular causes was 1.64 (95% CI, 0.98 to 2.74; P=0.06).”
- “[...] Patients and providers should consider the potential for serious adverse cardiovascular effects of treatment with rosiglitazone for type 2 diabetes.”

Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med. 2007 Jun 14;356(24):2457-71.

The mother of all meta-analyses in diabetes research



German Diabetes Center

P People with Type 2 Diabetes
I Rosiglitazone
C No rosiglitazone
O Number of cardiovascular deaths
S 42 RCTs



Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes

Steven E. Nissen, M.D., and Kathy Wolski, M.P.H.

- Number of citations (Google Scholar, June 2025): >6,400
- This analysis triggered the FDA to demand large cardiovascular outcome trials for new glucose-lowering drugs and paved the way for SGLT-2 inhibitors and GLP-1 receptor agonists to change (not only) type 2 diabetes treatment fundamentally.

Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med.* 2007 Jun 14;356(24):2457-71.

The mother of all meta-analyses in diabetes research

3 example trials from Nissen/Wolski

Study	Rosiglitazone		Control	
	#Events	#NonEvents	#Events	#NonEvents
49653/020	0	391	0	207
49653/082	1	211	0	107
ADOPT	2	1454	5	2890



Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes

Steven E. Nissen, M.D., and Kathy Wolski, M.P.H.

... pointing to some challenges: Double-zero trials, single-zero trials, unbalanced groups, heterogeneity

Standard analysis I: Inverse-variance methods

- Compute an effect estimate (e.g., the odds ratio) from each single study
- Compute the meta-analytic effect estimate by weighted linear regression (with the inverse estimation variance of the single study effect estimate as the weight)

Standard analysis I:

Inverse-variance methods



German Diabetes Center

Study	Rosiglitazone		Control		Log(Odds Ratio)	IV weight
	#E	#NE	#E	#NE		
49653/020	0	391	0	207	--	--
49653/082	1	211	0	107	--	--
ADOPT	2	1454	5	2890	-0.229	1.426

Challenges:

- „Two-step“-model: weights are assumed to be known (but are estimated)
- Empty cells lead to undefined weights

Standard analysis I: Inverse-variance methods

- “Many trials had few cardiovascular events, so the odds ratios and 95% confidence intervals were calculated with the use of the Peto method.”
- „Trials in which patients had no adverse cardiovascular events in either group were excluded from analyses.“



Effect of Rosiglitazone on the Risk of Myocardial Infarction
and Death from Cardiovascular Causes

Steven E. Nissen, M.D., and Kathy Wolski, M.P.H.

Standard analysis II:

Binary regression for correlated data

- Use two outcome proportions (one for treatment, one for control), a binary covariate of treatment and take into account that the two proportions come from the same study (are correlated)
- Effect estimator is determined by link function
- GLMM, GEE, Stratified analysis, ...

Standard analysis II:

Binary regression for correlated data

Study	Treatment	$p_E = \frac{\#E}{\#E + \#NE}$
49653/020	Rosiglitazone	0/391
49653/020	Control	0/207
49653/082	Rosiglitazone	1/212
49653/082	Control	0/107
ADOPT	Rosiglitazone	2/1456
ADOPT	Control	5/2895

Standard analysis II:

Binary regression for correlated data

Challenges:

- Which effects (intercept, treatment) should be considered random/fixed?
- Parameter estimation by ML needs numerically challenging methods or approximations

New analysis I: Multinomial models

New analysis I: Multinomial models

Study	(#ET, #NET, #EC, #NEC)
49653/020	(0, 391, 0, 207)
49653/082	(1, 211, 0, 107)
ADOPT	(2, 1454, 5, 2890)

- Each study constitutes a single observation from a (generalized) multinomial distribution
- For meta-analysis: Estimate marginal probabilities ($p_{ET}, p_{NET}, p_{EC}, p_{NEC}$) for each cell and combine them to achieve your preferred effect estimate (OR, RR, RD, ...)
- No need to fix the effect estimate in advance
- Estimate standard errors by multivariate delta method

New analysis I: Multinomial models (MN)

- Each study constitutes a single observation from a **standard multinomial (MN)** distribution
- Exact equivalence to the collapsed table

New analysis I: Multinomial models (MN)

Density (MN)

$$f(y_{ET}, \dots, y_{NEC}, p_{ET}, \dots, p_{NEC}) = \frac{\Gamma(\sum_k y_k + 1)}{\prod_k \Gamma(y_k + 1)} \prod_k p_k^{y_k}$$

with $k \in (ET, NET, EC, NEC)$, $\Gamma(*)$ the Gamma function,

and the y_k the absolute numbers from the study's fourfold table

	<i>E</i>	<i>NE</i>
<i>T</i>	y_{ET}	y_{NET}
<i>C</i>	y_{EC}	y_{NEC}

New analysis I: Multinomial models (DM)

- Each study constitutes a single observation from a **Dirichlet multinomial (DM)** distribution
- The marginal probabilities ($p_{ET}, p_{NET}, p_{EC}, p_{NEC}$) are drawn from a Dirichlet distribution with parameters $\alpha_{ET}, \alpha_{NET}, \alpha_{EC}, \alpha_{NEC}$
- **One** additional parameter ($\alpha_0 = \alpha_{ET} + \alpha_{NET} + \alpha_{EC} + \alpha_{NEC}$) as compared to the MN distribution to model heterogeneity between studies

New analysis I: Multinomial models (DM)

Density (DM)

$$f(y_{ET}, \dots, y_{NEC}, \alpha_{ET}, \dots, \alpha_{NEC}) = \frac{\Gamma(\alpha_0) \Gamma(\sum_k y_k + 1)}{\Gamma(y_k + \alpha_0)} \prod_k \frac{\Gamma(y_k + \alpha_k)}{\Gamma(\alpha_0) \Gamma(y_k + 1)}$$

with $k \in (ET, NET, EC, NEC)$, $\alpha_0 = \sum_k \alpha_k$, $\Gamma(*)$ the Gamma function,
and the y_k the absolute numbers from the study's fourfold table

New analysis I:

Multinomial models (DNM)



German Diabetes Center

- Each study constitutes a single observation from a **Dirichlet negative multinomial (DNM)** distribution
- The marginal probabilities $(p_{ET}, p_{NET}, p_{EC}, p_{NEC})$ are drawn from a Dirichlet distribution with parameters $\alpha_{ET}, \alpha_{NET}, \alpha_{EC}, \alpha_{NEC}$
- The event numbers $(y_{ET}, y_{NET}, y_{EC}, y_{NEC})$ are drawn from a negative multinomial distribution with additional parameter y_0
- **Two** additional parameters $(\alpha_0 = \alpha_{ET} + \alpha_{NET} + \alpha_{EC} + \alpha_{NEC}, y_0)$ as compared to the MN distribution to model heterogeneity between studies

Density (DNM)

$$f(y_{ET}, \dots, y_{NEC}, \alpha_{ET}, \dots, \alpha_{NEC}) = \frac{B(y_0 + \sum_k y_k, \alpha_0 + \sum_k \alpha_k)}{B(y_0, \alpha_0)} \prod_k \frac{\Gamma(y_k + \alpha_k)}{x_k! \Gamma(\alpha_k + 1)}$$

with $k \in (ET, NET, EC, NEC)$, $\alpha_0 = \sum_k \alpha_k$,

$B(*,*)$ the Beta function, $\Gamma(*)$ the Gamma function,

and the y_k the absolute numbers from the study's fourfold table

Additional insights from econometrics!

There is an exact equivalence between the (generalized) multinomial models and the class of Panel Count Data models from econometrics [Hausman1984]

Multinomial model
Standard Multinomial (MN)
Negative multinomial (NM)
Dirichlet multinomial (DM)
Dirichlet negative multinomial (DNM)

Additional insights from econometrics!

There is an exact equivalence between the (generalized) multinomial models and the class of Panel count data models from econometrics [Hausman1984]

Multinomial model	Panel count data model
Standard Multinomial (MN)	FE-Poisson
Negative multinomial (NM)	RE-Poisson
Dirichlet multinomial (DM)	FE-Negbin
Dirichlet negative multinomial (DNM)	RE-Negbin
FE: Fixed Effects, RE: Random Effects	

New analysis II:

Panel count data models



German Diabetes Center

- In econometrics, a „panel“ is a short time series with correlated observations
- **Here:** The single study equals the panel
- Each panel has exactly four observations (= the four cells in the study's four-fold table)

New analysis II: Panel count data models

$$E(y_{it}|x_{it}, \theta_i) = \mu_{it} = \theta_i \lambda_{it} = \theta_i \exp(x'_{it}\beta)$$

with $i = 1, \dots, n$ indexing the studies and $t = 1, \dots, 4$ the cells,

y_{it} the outcome (number of events or non-events in the study's fourfold table)

θ_i a study-specific effect which acts multiplicatively on the outcome

x'_{it} a 4 x 1 vector (with covariates 1, Treatment type, Event type, Treatment type x Event type)

β the respective parameters

New analysis II:

Panel count data models



German Diabetes Center

Study	Y (= #Events)	Treatment type	Event type	Treatment type x Event type
49653/020	0	1 (Rosiglitazone)	1	1
49653/020	391	1 (Rosiglitazone)	0	0
49653/020	0	0 (Control)	1	0
49653/020	207	0 (Control)	0	0
49653/082	1	1 (Rosiglitazone)	1	1
49653/082	211	1 (Rosiglitazone)	0	0
49653/082	0	0 (Control)	1	0
49653/082	107	0 (Control)	0	0
ADOPT	2	1 (Rosiglitazone)	1	1
ADOPT	1454	1 (Rosiglitazone)	0	0
ADOPT	5	0 (Control)	1	0
ADOPT	2890	0 (Control)	0	0

New analysis II: Panel count data models

In principle, the study-specific intercepts θ_i could be fitted ...

- ... like a categorical covariate with $n - 1$ parameters or
- ... by assuming a distribution for θ_i

But what we actually do is to condition them out of the likelihood
(where conditioning is on the sum of individual observations in a study $\sum_{t=1}^4 y_{it}$)

New analysis II: Panel count data models

Conditioning out the θ_i means ...

- ... that no assumptions on its distribution must be made
- ... that in the FE models they are allowed to be correlated with the regressors x ,
- ... that the resulting likelihood is also a multinomial likelihood
- ... that no study-specific effects can be fitted (no Meta-regression)

New analysis II: Panel count data models

Multinomial model	Panel count data model	Assumptions on the distribution of ...	
		... y	... study-specific effects
MN	FE-Poisson	Poisson	None
NM	RE-Poisson	Poisson	$\theta_i \sim \text{gamma}(\delta, \delta)$
DM	FE-Negbin	Negative binomial with parameter ν_i	None
DNM	RE-Negbin	Negative binomial with parameter ν_i	$(1 + \theta_i/\nu_i)^{-1} \sim B(a, b)$

FE: Fixed Effects, RE: Random Effects

New analysis II:

Panel count data models



German Diabetes Center

- „Fixed“ and „random“ here mean entirely different things than we (at least in biostatistics) are used to
- “[...] the term “fixed” as used here indicates that the term does not vary over time [here: across cells], not that it is nonstochastic.” [Greene2003]

New analysis II:

Panel count data models



German Diabetes Center

- The FE-models are the **less restrictive** models (as compared to the parallel RE-model)
- The actual effect of treatment is measured by the interaction effect Treatment type x Event type ($\beta_{Tt \times Et} = \text{LogOR}$).
- Conjugation relations assure closed-form likelihoods also in the RE-cases

New analysis II: Panel count data models

- In the FE-Poisson model the Conditional ML estimator coincides with ...
 - ... the collapsed table (because MN = FE-Poisson)
 - ... the ML estimator (yielding the resp. good properties)
 - ... the moment-based/quasi-likelihood estimator with the only moment condition [Wooldridge1999]

$$E(y_{it}|x_{it}, \theta_i) = \theta_i \exp(x'_{it}\beta)$$

That means, we don't even need the Poisson assumption for the outcome and get a „semi-parametric“ estimator

New analysis II: Panel count data models

- “Overall, this robustness property is a very useful aspect of the FE-Poisson model. Thus one does not need to worry about overdispersion, or other expressions of “non-Poissonness.” [Winkelmann2008]
- Woolridge [Woolridge1999] recommends to compute robust standard errors

[Winkelmann2008] Winkelmann R. *Econometric Analysis of Count Data*, Fifth edition. 2008. Springer-Verlag Berlin Heidelberg. ISBN 978-3-540-77648-2, e-ISBN 978-3-540-78389-3, DOI: 10.1007/978-3-540-78389-3.

[Woolridge1999] Wooldridge JM. Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics* 90 (1999) 77-97.

Multinomial models

SAS

MN `proc logistic ...; model ... / link=glogit; freq y; run;`
`proc fmm ...; model ySuccessTreatment yFailureTreatment
 ySuccessControl yFailureControl= / dist=multinomial;`

DM `proc fmm ...; model ySuccessTreatment yFailureTreatment
 ySuccessControl yFailureControl=/dist=dirichletmultinomial;`

DNM `proc nlmixed with hand-coded likelihood`

R package MGLM (MN, DM, DNM)

Zhang Y, Zhou H. MGLM: Multivariate Response Generalized Linear Models. <https://cran.r-project.org/package=MGLM>.

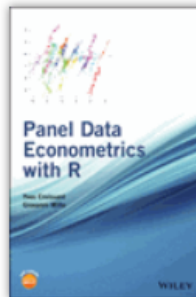
Zhang Y, Zhou H, Zhou J, Sun W. Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*. 2017;26(1):1–13.

Panel count data models

SAS

```
proc countreg groupid=Study;  
  model Y=treatmenttype eventtype treatmenttype*eventtype /  
    errorcomp=fixed dist=poisson /* (FE-Poisson) */  
    errorcomp=fixed dist=negbin /* (FE-Negbin) */  
    errorcomp=random dist=negbin /* (RE-Negbin) */;  
run;
```

R package pglm



Panel Data Econometrics with R

Editor(s): Yves Croissant, Giovanni Millo

First published: 10 August 2018

Print ISBN: 9781118949160 | Online ISBN: 9781119504641 |

DOI: 10.1002/9781119504641

© 2019 John Wiley & Sons, Ltd

Collapsed table from Nissen/Wolski:

	Death from cardiovascular causes		
	Yes	No	
Rosiglitazone	39	15,517	15,556
Control	22	12,255	12,277

Odds Ratio from collapsed table

(= MN = FE-Poisson):

1.40 (95% CI, 0.83 to 2.36; P=0.20)



Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes

Steven E. Nissen, M.D., and Kathy Wolski, M.P.H.

“In the rosiglitazone group, as compared with the control group, the odds ratio [...] for death from cardiovascular causes was 1.64 (95% CI, 0.98 to 2.74; P=0.06).”

Results

Model	Odds Ratio (95% CI; p)
Nissen/Wolski	1.64 (0.98 to 2.74; 0.06)
MN/FE-Poisson/Collapsed table	1.40 (0.83 to 2.36; 0.20)
MN/FE-Poisson/Collapsed table (with robust SEs)	1.40 (1.01 to 1.95; 0.04)
DM/FE-Negbin	1.88 (0.82 to 4.27; 0.13)
DNM/RE-Negbin	1.88 (0.33 to 3.42; 0.26)



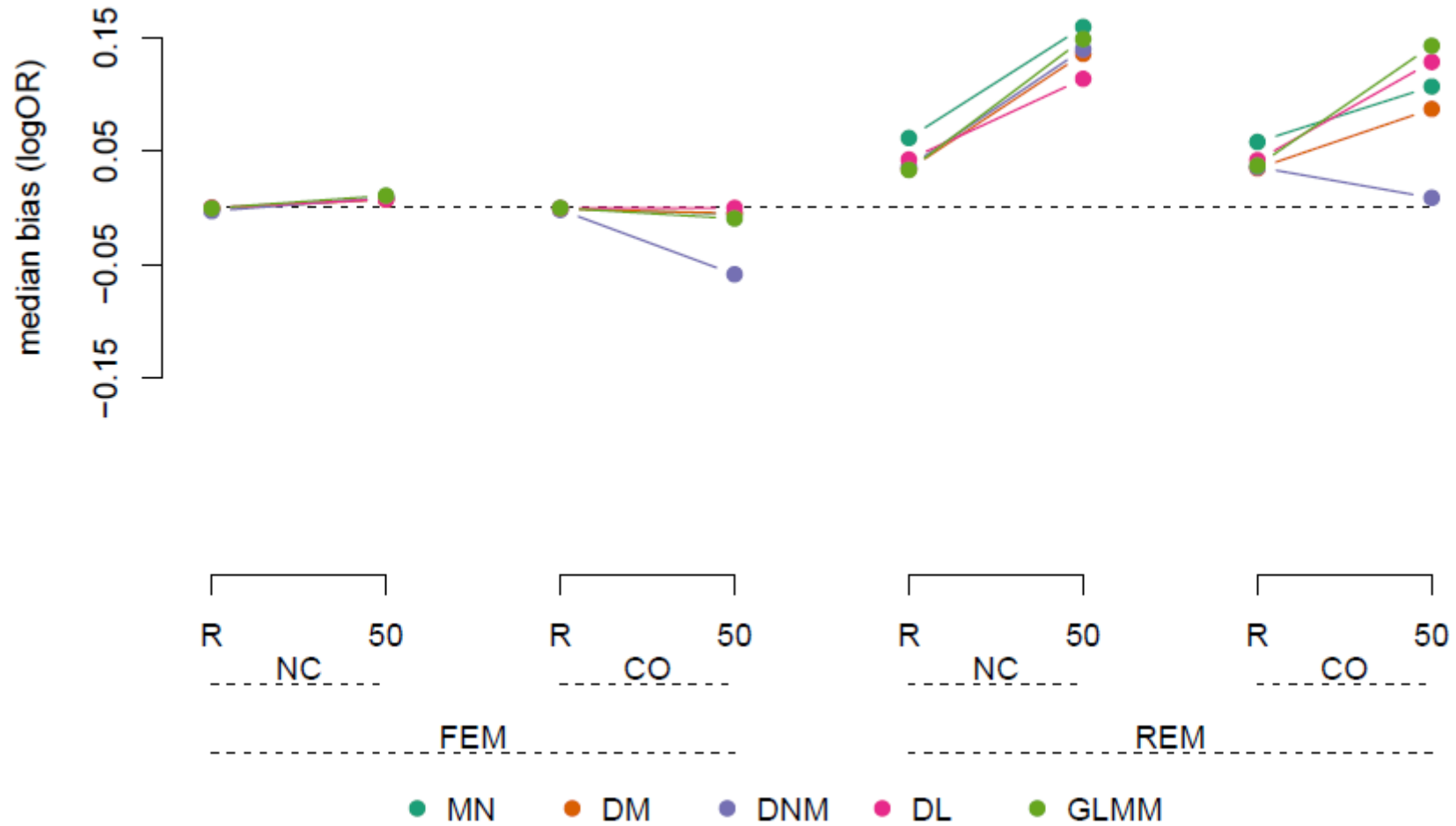
Effect of Rosiglitazone on the Risk of Myocardial Infarction
and Death from Cardiovascular Causes

Steven E. Nissen, M.D., and Kathy Wolski, M.P.H.

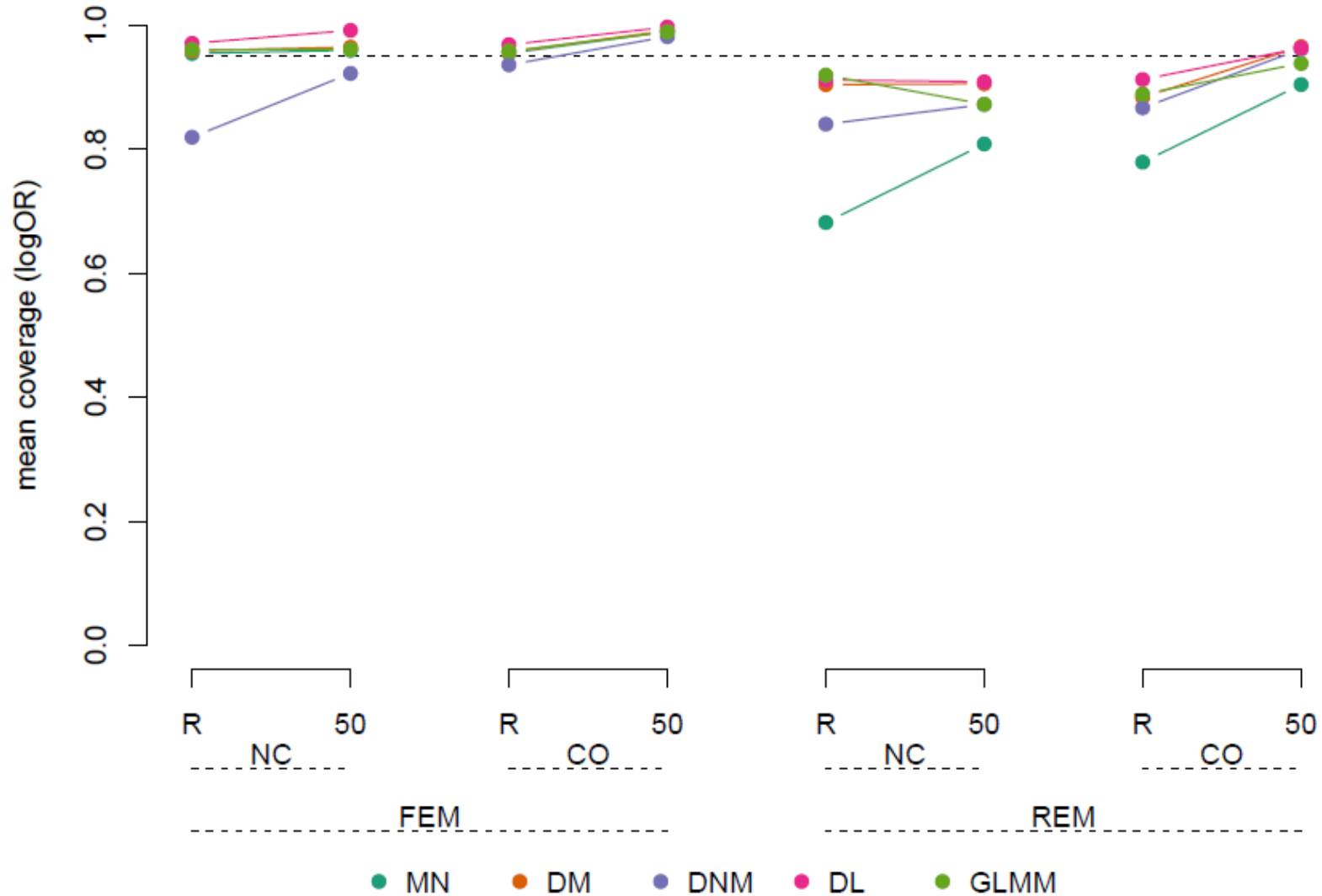
(Preliminary) Simulation

- Simulation setting from Kuss(2015)
- IV-MA is the true model
- **Methods:** MN, DM, DNM, DL (DerSimonianLaird), GLMM
- **Outcomes:** Median bias, Coverage, Convergence for the LogOR
- 10,000 generated data sets

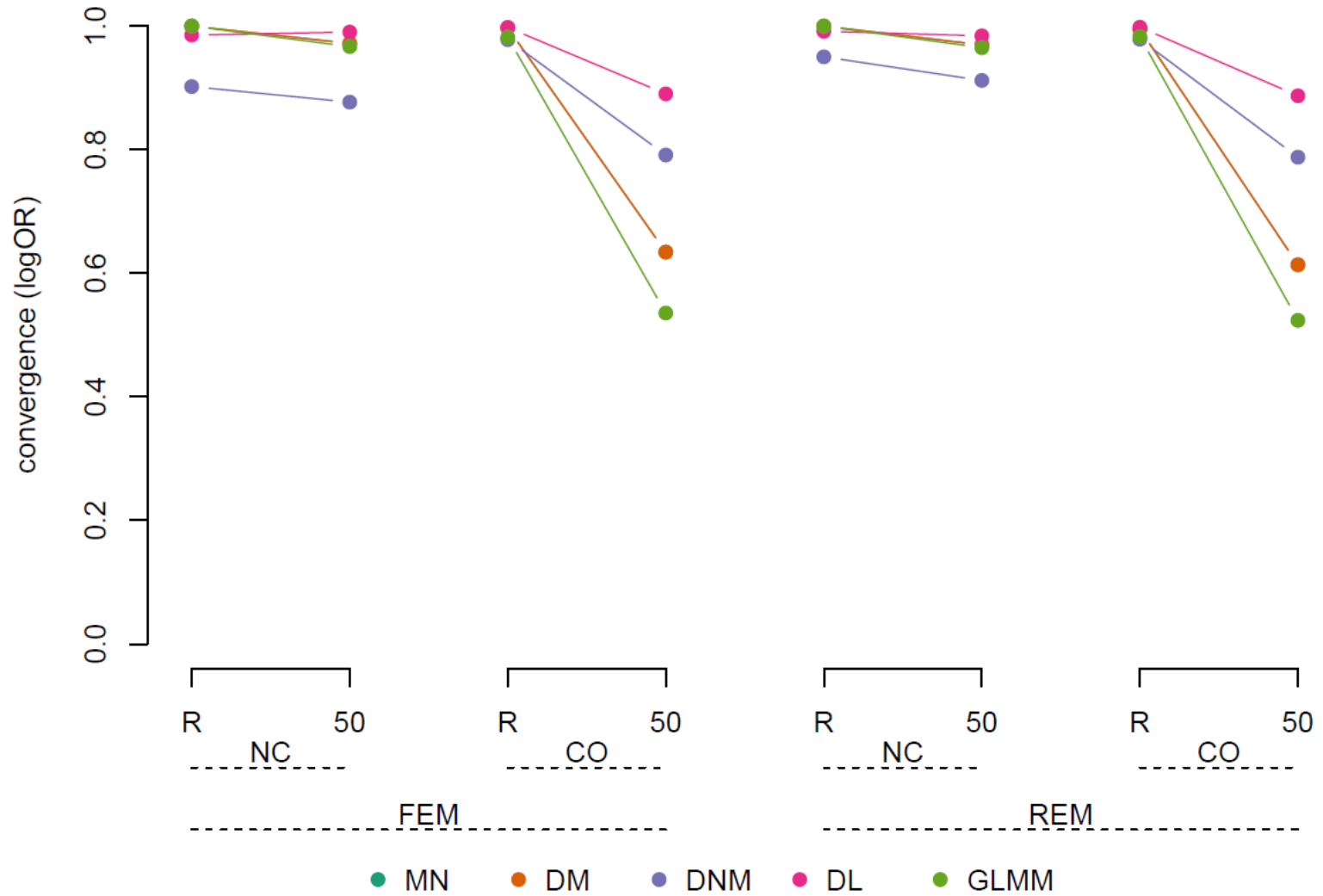
(Preliminary) Simulation



(Preliminary) Simulation



(Preliminary) Simulation



We propose to use (generalized) multinomial distributions/Panel count data models for the meta-analysis of binary outcomes

Advantages (from the multinomial view)

- Closed-form likelihood with 1 or even 2 additional parameters accounting for heterogeneity
- One-step approach
- Generic approach, no need to pre-select a single effect measure
- Natural handling of single-zero- or double-zero studies

Discussion

Advantages (from the Panel count data view)

- Conditioning out the individual study effects saves from having to make assumptions for those

Limitations (from the multinomial view)

- Parameters of interest (OR, RR, RD) are not model parameters, but (non-linear) combinations of them, which makes computation of SEs complicated

Limitations (from the Panel count data view)

- Between-study effects are not identified → Meta-regression not possible
- Only information from within, but not between studies is used. This may yield larger SEs

- Simulation evidence is needed
- Translation of heterogeneity in IV-MA (τ^2) to multinomial parameters
- Generalization to network meta-analysis or meta-analysis of diagnostic trials
- There is a Generalized Dirichlet multinomial distribution (GDM) with a more general dependence structure, but without a parallel PCD-model.

Connor RJ, Mosimann JE. Concepts of Independence for Proportions With a Generalization of the Dirichlet Distribution. *Journal of the American Statistical Association*, 1969, 64, 194–206.

Bouguila N. Clustering of Count Data Using Generalized Dirichlet Multinomial Distributions. *IEEE Transactions on Knowledge and Data Engineering*, Apr. 2008, pp. 462-474, vol. 20.

- Explore connections to other distributions
(DNM = RE-Negbin = Multivariate Generalized Waring distribution)
Xekalaki, E. (1986). The multivariate generalized Waring distribution. *Communications in Statistics-Theory and Methods*, 15, 1047–1064.
- Check the vulnerability against Simpson's paradoxon
Rücker G, Schumacher M. Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. *BMC Med Res Methodol*. 2008 May 30;8:34

A provocative thesis

Forget about all meta-analytic methods that use the standard principles of inverse-variance or binary regression for correlated data!

The collapsed table (= MN model = FE-Poisson model) is enough!

A provocative thesis

- In particular, aggregating all studies in a single four-fold table ...
 - ... is not „naive pooling“ [Chuang-Stein2011, Röver2025, Schwarzer2025],
 - ... is not „simple unstratified lumping“ [Wittes2015], and
 - ... does not „explicitly ignore that data were collected from several studies, thus assuming that the underlying risk of an event is constant across trials.“ [Kuss2013]

Chuang-Stein C, Beltangady M. Reporting cumulative proportion of subjects with an adverse event based on data from multiple studies. Pharm Stat. 2011 Jan-Feb;10(1):3-7.

Röver C. An introduction to meta-analysis using the bayesmeta package. <https://cran.r-project.org/web/packages/bayesmeta/vignettes/bayesmeta.html>. 2025

Schwarzer G. Package 'meta'. <https://cran.r-project.org/web/packages/meta/meta.pdf>. 2025

Wittes J, Crowe B, Chuang-Stein C, Guettner A, Hall D, Jiang Q, Odenheimer D, Xia HA, Kramer J. The FDA's Final Rule on Expedited Safety Reporting: Statistical Considerations. Stat Biopharm Res. 2015 Jul 3;7(3):174-190.

Kuss O. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. Stat Med. 2015 Mar 30;34(7):1097-116.

A provocative thesis

- The collapsed table (if seen as a FE-Poisson PCD model) is rather the most robust (= imposing the minimum number of assumptions) method available that ...
 - ... accounts fully for all differences between studies
 - ... makes no assumptions the study-specific intercepts
 - ... does not need the Poisson (or any other distribution) assumption for the outcome, but only the exponential form of the linear (mean) predictor

Thank you

Mach mit!

