Krause, Joscha; Burgard, Jan Pablo; Morales, Domingo

ℓ2-penalized approximate likelihood inference in logit mixed models for regional prevalence estimation under covariate rank-deficiency

Metrika

Provided in Cooperation with:
Springer Nature

Suggested Citation: Krause, Joscha; Burgard, Jan Pablo; Morales, Domingo (2021) : ℓ2-penalized approximate likelihood inference in logit mixed models for regional prevalence estimation under covariate rank-deficiency, Metrika, ISSN 1435-926X, Springer, Berlin, Heidelberg, Vol. 85, Iss. 4, pp. 459-489,
https://doi.org/10.1007/s00184-021-00837-y

Metrika (2022) 85:459–489
https://doi.org/10.1007/s00184-021-00837-y

ℓ_2-penalized approximate likelihood inference in logit mixed models for regional prevalence estimation under covariate rank-deficiency

Joscha Krause[1] · **Jan Pablo Burgard**[1] · **Domingo Morales**[2]

Received: 1 October 2020 / Accepted: 28 July 2021 / Published online: 17 August 2021

Abstract
Regional prevalence estimation requires the use of suitable statistical methods on epidemiologic data with substantial local detail. Small area estimation with medical treatment records as covariates marks a promising combination for this purpose. However, medical routine data often has strong internal correlation due to diagnosis-related grouping in the records. Depending on the strength of the correlation, the space spanned by the covariates can become rank-deficient. In this case, prevalence estimates suffer from unacceptable uncertainty as the individual contributions of the covariates to the model cannot be identified properly. We propose an area-level logit mixed model for regional prevalence estimation with a new fitting algorithm to solve this problem. We extend the Laplace approximation to the log-likelihood by an ℓ_2-penalty in order to stabilize the estimation process in the presence of covariate rank-deficiency. Empirical best predictors under the model and a parametric bootstrap for mean squared error estimation are presented. A Monte Carlo simulation study is conducted to evaluate the properties of our methodology in a controlled environment. We further provide an empirical application where the district-level prevalence of multiple sclerosis in Germany is estimated using health insurance records.

Keywords Generalized linear mixed models · Laplace approximation · Multiple sclerosis · Prevalence mapping · Small area estimation

✉ Joscha Krause
krause@uni-trier.de

[1] Department of Economic and Social Statistics, Trier University, Universitätsring 15, 54296 Trier, Germany

[2] Operations Research Center, University Miguel Hernández de Elche, Elche, Spain

1 Introduction

Regional prevalence estimation is an essential element of modern epidemiologic research (Branscum et al. 2008; Stern 2014; Burgard et al. 2019). Policymakers and health care providers need reliable information on regional disease distributions to plan comprehensive health programs. Depending on the disease of interest, corresponding figures may not be recorded in registries and must be estimated from survey data instead. However, national health surveys often lack in sufficient local observations due to limited resources. As a result, regional prevalence estimates based on survey data can be subject to unacceptable uncertainty due to large sampling variances. Small area estimation (SAE) solves this problem by linking a response variable of interest to statistically related covariates by means of a suitable statistical model. The observations from multiple regions are combined and jointly used for model parameter estimation. Regional prevalence estimates are obtained via model prediction, which allows for an increase in the effective sample size relative to classical direct estimation. See Rao and Molina (2015) for an overview on SAE.

In practice, the efficiency advantage of SAE methods over direct estimators is mainly determined by two aspects: (i) finding a suitable model type to describe the response variable, and (ii) having covariate data with explanatory power. Regarding the first aspect, since regional prevalence is usually stated as proportion (number of sick persons divided by the total number of persons), binomial, Poisson or negative binomial mixed models are canonical choices. The binomial-logit approach has been used for regional proportion estimation in the past, for instance by Molina et al. (2007), Ghosh et al. (2009), Chen and Lahiri (2012), Erciulescu and Fuller (2013), López-Vizcaíno et al. (2013), López-Vizcaíno et al. (2015), Burgard (2015), Militino et al. (2015), Chambers et al. (2016), Hobza and Morales (2016), Liu and Lahiri (2017) and Hobza et al. (2018). The Poisson or negative binomial mixed models were applied to estimate small area counts or proportions by Berg (2010), Chambers et al. (2014), Dreassi et al. (2014), Tzavidis et al. (2015) and Boubeta et al. (2016, 2017), among others. Marino et al. (2019) propose a semiparametric approach allowing for a flexible random effects structure in unit-level models. Ranalli et al. (2018) introduced benchmarking for logistic unit-level. Concerning the second aspect, medical routine data provided by official statistics or health insurance companies have been found to be promising data bases for regional prevalence estimation. Exemplary applications were provided by Tamayo et al. (2016), Burgard et al. (2019), and Breitkreutz et al. (2019).

However, using medical routine data as covariates can be problematic, especially within logit mixed models. Medical treatment frequencies are typically recorded and coded into diagnosis groups, for instance on ICD-3 level (World Health Organization 2018). This context-related segmentation can induce strong correlation between medical treatment frequencies for diseases that are closely related in terms of comorbidity, such as diabetes and hypertension (Long and Dagogo-Jack 2011). If two or more diagnoses from the auxiliary data set are strongly correlated, the space spanned by the covariates can become rank-deficient. In that case, it is not possible to accurately separate the individual contributions of the covariates to the description of the response variable. Model parameter estimates suffer from high variance and model predic-

tions for regional prevalence are not reliable. This is particularly problematic for logit mixed models, as model parameter estimation already relies on approximate inference in the absence of rank-deficiency (Breslow and Clayton 1993). The respective likelihood integral does not have a closed-form solution, which requires techniques like the Laplace approximation to find a proper substitute as objective function. Therefore, when approximate inference is to be performed on a rank-deficient covariate space, methodological adjustments are required to allow for reliable results.

In this paper, we propose a modification to the maximum likelihood Laplace (ML-Laplace) algorithm for model parameter estimation (e.g. Demidenko 2013; Hobza et al. 2018) in a logit mixed model under covariate rank-deficiency. We draw from theoretical insights on ridge regression (Hoerl and Kennard 1970) and extend the Laplace approximation to the log-likelihood function by the squared ℓ_2-norm of the regression parameters (ℓ_2-penalty). This adjustment reduces the variance of model parameter estimates considerably and improves approximate inference in the presence of strong covariate correlation. To the best of our knowledge, ℓ_2-penalization has only been studied for standard ML estimation in fixed effect logit models, for instance by Schaefer et al. (1984), Cessie and Houwelingen (1992), and Pereira et al. (2016). We are not aware of corresponding studies for logit mixed models based on ML-Laplace estimation, especially not in the context of SAE.

An area-level binomial logit mixed model for regional prevalence estimation is presented. Following Jiang and Lahiri (2001) and Jiang (2003), we derive empirical best predictors (EBPs) under the model and present a parametric bootstrap estimator for their mean squared error (MSE). Thereafter, we state the Laplace approximation to the log-likelihood function and demonstrate ℓ_2-penalized approximate likelihood (ℓ_2-PAML) estimation of the model parameters. We further show how the tuning parameter for the ℓ_2-penalty can be chosen in practice. A Monte Carlo simulation study is conducted to study the behavior of ℓ_2-PAML estimation under different degrees of covariate correlation. And finally, the proposed methodology is applied to regional prevalence estimation in Germany. We use health insurance records of the German Public Health Insurance Company (AOK) and inpatient diagnosis frequencies of the Diagnosis-Related Group Statistics (DRG-Statistics) to estimate district-level multiple sclerosis prevalence.

The remainder of the paper is organized as follows. In Sect. 2, we present the model and its EBP. We further address MSE estimation. In Sect. 3, we present the Laplace approximation and the technical details for ℓ_2-PAML. Section 4 contains a Monte Carlo simulation study. Section 5 covers the application to regional prevalence estimation. Section 6 closes with some conclusive remarks.

2 Model

2.1 Formulation

For the subsequent derivation, we rely on model-based inference in a finite population setting. Let U be a finite population of size $|U| = N$. Suppose that U is partitioned into D domains U_d of size $|U_d| = N_d$. That is to say, $U = \cup_{d=1}^{D} U_d$, $U_{d_1} \cap U_{d_2} = \emptyset$,

$d_1 \neq d_2$, and $\sum_{d=1}^{D} N_d = N$. Let S be a sample of size $|S| = n$ that is drawn from U. For simplicity, assume the sampling scheme is such that there are domain-specific subsamples S_d of size $|S_d| = n_d$ with fixed $n_d > 0$ for all $d = 1, ..., D$. Thus, we have $S = \cup_{d=1}^{D} S_d$ and $\sum_{d=1}^{D} n_d = n$. Let y be a dichotomous response variable with potential outcomes $\{0, 1\}$. Denote the realization of y for some individual $i \in U_d$ by y_{id}. Note that we use the same symbol for a random variable and its realizations in order to avoid overloading the notation. Define $y_d = \sum_{i \in s_d} y_{id}$ as the sample total (count) of y in domain U_d. Let $x = \{x_1, ..., x_p\}$ be a set of covariates statistically related to y. Denote $\boldsymbol{x}_d$ as a $1 \times p$ vector of aggregated (domain-level) realizations of x. Suppose that corresponding information is retrieved from administrative records and not calculated from the sample S. In what follows, we present an area-level logit mixed model for estimating the domain totals $Y_d = \sum_{i \in U_d} y_{id}$ or proportions $p_d = Y_d / N_d$ of the response variable. Let us consider a set of random effects such that $\{v_d : d = 1, \ldots, D\}$ are independent and identically distributed according to $v_d \sim N(0, 1)$. In matrix notation, we have normally distributed random effects

$$\boldsymbol{v} = \underset{1 \leq d \leq D}{\text{col}}(v_d) \sim N_D(\boldsymbol{0}, \boldsymbol{I}_D) \tag{1}$$

and, hence, their probability density function (PDF) is stated as

$$f_v(\boldsymbol{v}) = (2\pi)^{-D/2} \exp\Big\{ -\frac{1}{2} \boldsymbol{v}'\boldsymbol{v} \Big\}. \tag{2}$$

The model assumes that the distribution of the response variable y_d, conditioned to the random effect v_d, is

$$y_d | v_d \sim \text{Bin}(n_d, p_d), \quad d = 1, \ldots, D, \tag{3}$$

and that the natural parameter fulfills

$$\eta_d = \log \frac{p_d}{1 - p_d} = \boldsymbol{x}_d \boldsymbol{\beta} + \phi v_d, \quad d = 1, \ldots, D, \tag{4}$$

where $\phi > 0$ is an standard deviation parameter, $\boldsymbol{\beta} = \text{col}_{1 \leq r \leq p}(\beta_r)$ is the vector of regression parameters and $\boldsymbol{x}_d = \text{col}'_{1 \leq r \leq p}(x_{dr})$. We complete the model definition by assuming that the domain-specific sample counts y_d are independent when conditioned on the random effects $\boldsymbol{v}$. Therefore, the conditional PDF of $\boldsymbol{y} = \text{col}_{1 \leq d \leq D}(y_d)$ given $\boldsymbol{v}$ is stated as

$$P(\boldsymbol{y}|\boldsymbol{v}) = \prod_{d=1}^{D} P(y_d|v_d), \quad P(y_d|\boldsymbol{v}) = P(y_d|v_d) = \binom{n_d}{y_d} p_d^{y_d} (1 - p_d)^{n_d - y_d}, \tag{5}$$

where

$$p_d = \frac{\exp\{\boldsymbol{x}_d \boldsymbol{\beta} + \phi v_d\}}{1 + \exp\{\boldsymbol{x}_d \boldsymbol{\beta} + \phi v_d\}} = \frac{\exp\{\eta_d\}}{1 + \exp\{\eta_d\}}, \quad 1 - p_d = \frac{1}{1 + \exp\{\eta_d\}}. \tag{6}$$

Further, the unconditional PDF of $\boldsymbol{y}$ is

$$P(\boldsymbol{y}) = \int_{R^D} P(\boldsymbol{y}|\boldsymbol{v}) f_v(\boldsymbol{v})\, d\boldsymbol{v} = \int_{R^D} \psi(\boldsymbol{y}, \boldsymbol{v})\, d\boldsymbol{v}, \tag{7}$$

with

$$\begin{aligned}\psi(\boldsymbol{y}, \boldsymbol{v}) &= (2\pi)^{-\frac{D}{2}} \exp\left\{\frac{-\boldsymbol{v}'\boldsymbol{v}}{2}\right\} \prod_{d=1}^{D} \frac{\binom{n_d}{y_d} \exp\{y_d(\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d)\}}{\left[1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\}\right]^{n_d}} \\ &= (2\pi)^{-\frac{D}{2}} \prod_{d=1}^{D} \binom{n_d}{y_d} \exp\left\{\frac{-\boldsymbol{v}'\boldsymbol{v}}{2}\right\} \exp\left\{\sum_{k=1}^{p} \Big(\sum_{d=1}^{D} y_d x_{dk}\Big)\beta_k + \phi \sum_{d=1}^{D} y_d v_d \right. \\ &\quad \left. - \sum_{d=1}^{D} n_d \log\left(1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\}\right)\right\}. \end{aligned} \tag{8}$$

2.2 Prediction

Hereafter, we obtain EBPs under the area-level logit mixed model introduced in Sect. 2.1. For this, we first derive best predictors (BPs) in a preliminary setting where all model parameters $\boldsymbol{\theta} := (\boldsymbol{\beta}', \phi)$ are assumed to be known. Then, the EBPs are obtained by replacing the full parameter vector $\boldsymbol{\theta}$ by consistent estimators $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{\beta}}', \hat{\phi})$. Note that calculating the EBP requires Monte Carlo integration over the random effect PDF, which can be computationally infeasible for some applications. Therefore, we also state two alternative predictors that do not rely on Monte Carlo integration and are easier to apply in practice. We start with the EBPs. Recall the definition of the conditional PDF $P(\boldsymbol{y}|\boldsymbol{v})$ from the last section. For the domain-specific component $P(y_d|v_d)$, we can write

$$\begin{aligned} P(y_d|v_d) &= \binom{n_d}{y_d} p_d^{y_d} (1 - p_d)^{n_d - y_d} = \frac{\binom{n_d}{y_d} \exp\{y_d(\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d)\}}{\left[1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\}\right]^{n_d}} \\ &= \exp\left\{\log\binom{n_d}{y_d} + y_d(\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d) - n_d \log\left[1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\}\right]\right\} \end{aligned} \tag{9}$$

The probability density function of $\boldsymbol{v}$ is

$$f(\boldsymbol{v}) = \prod_{d=1}^{D} f(v_d), \quad f(v_d) = (2\pi)^{-1/2} \exp\Big\{-\frac{1}{2} v_d^2\Big\}. \tag{10}$$

The BP of $p_d = p_d(\boldsymbol{\theta}, v_d)$ is given by the conditional expectation $\hat{p}_d(\boldsymbol{\theta}) = E_\theta[p_d|\boldsymbol{y}]$. Due to the conditional independence of the response realizations given the random effects, we have $E_\theta[p_d|\boldsymbol{y}] = E_\theta[p_d|y_d]$ and

$$E_\theta[p_d|y_d] = \frac{\int_R \frac{\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}}{1+\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}} P(y_d|v_d) f(v_d)\, dv_d}{\int_R P(y_d|v_d) f(v_d)\, dv_d} = \frac{\mathcal{N}_d(y_d,\boldsymbol{\theta})}{\mathcal{D}_d(y_d,\boldsymbol{\theta})} = \frac{N_d(y_d,\boldsymbol{\theta})}{D_d(y_d,\boldsymbol{\theta})}, \tag{11}$$

where $\mathcal{N}_d = \mathcal{N}_d(y_d,\boldsymbol{\theta})$, $\mathcal{D}_d = \mathcal{D}_d(y_d,\boldsymbol{\theta})$, $N_d = N_d(y_d,\boldsymbol{\theta})$ and $D_d = D_d(y_d,\boldsymbol{\theta})$ are functions of the model parameters and the domain-specific sample counts. They are stated as follows:

$$\begin{aligned}
\mathcal{N}_d &= \int_R \frac{\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}}{1+\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}} \exp\left\{\log\binom{n_d}{y_d} + y_d\boldsymbol{x}_d\boldsymbol{\beta} + \phi y_d v_d \right.\\
&\quad \left. - n_d \log\left[1+\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}\right]\right\} f(v_d)\, dv_d,\\
\mathcal{D}_d &= \int_R \exp\left\{\log\binom{n_d}{y_d} + y_d\boldsymbol{x}_d\boldsymbol{\beta} + \phi y_d v_d - n_d \log\left[1+\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}\right]\right\} f(v_d)\, dv_d,\\
N_d &= \int_R \frac{\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}}{1+\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}} \exp\left\{\phi y_d v_d - n_d \log\left[1+\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}\right]\right\} f(v_d)\, dv_d,\\
D_d &= \int_R \exp\left\{\phi y_d v_d - n_d \log\left[1+\exp\{\boldsymbol{x}_d\boldsymbol{\beta}+\phi v_d\}\right]\right\} f(v_d)\, dv_d.
\end{aligned}$$

We can conclude that the EBP of p_d is $\hat{p}_d(\hat{\boldsymbol{\theta}})$. However, its quantification requires integration over the random effect PDF $f(v_d)$. As the logit mixed model belongs to the family of generalized linear mixed models, this cannot be performed analytically. Instead, we apply Monte Carlo integration and approximate the EBP as follows:

1. Estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\phi})$.
2. For $k = 1, \ldots, K$, generate $v_d^{(k)}$ i.i.d. $N(0, 1)$ and $v_d^{(K+k)} = -v_d^{(k)}$.
3. Calculate $\hat{p}_d(\hat{\boldsymbol{\theta}}) = \hat{N}_d / \hat{D}_d$, where

$$\begin{aligned}
\hat{N}_d &= \frac{1}{2K}\sum_{k=1}^{2K}\left\{\frac{\exp\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}+\hat{\phi}v_d^{(k)}\}}{1+\exp\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}+\hat{\phi}v_d^{(k)}\}} \exp\left\{\hat{\phi}y_d v_d^{(k)} - n_d \log\left[1+\exp\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}+\hat{\phi}v_d^{(k)}\}\right]\right\}\right\},\\
\hat{D}_d &= \frac{1}{2K}\sum_{k=1}^{2K}\exp\left\{\hat{\phi}y_d v_d^{(k)} - n_d \log\left[1+\exp\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}+\hat{\phi}v_d^{(k)}\}\right]\right\}.
\end{aligned}$$

The EBP of p_d can be used to obtain the predictor $\hat{Y}_d = N_d\,\hat{p}(\hat{\boldsymbol{\theta}})$ of the domain total Y_d.

We now state two alternative predictors that do not rely on Monte Carlo integration. The first is a synthetic predictor. It is characterized by a regression-synthetic prediction from the area-level logit mixed model without considering the random effect. On that note, the synthetic predictor of p_d is obtained according to

$$\tilde{p}_d^{syn} = \frac{\exp\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}\}}{1+\exp\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}\}}, \tag{12}$$

which constitutes the synthetic predictor $\tilde{Y}_d^{syn} = N_d\,\tilde{p}_d^{syn}$ for Y_d. The plug-in predictor is obtained along the same lines, but includes the random effects v_d as well as the

variance parameter ϕ. For the prediction of p_d, we have

$$\tilde{p}_d^{plug} = \frac{\exp\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}} + \hat{\phi}\hat{v}_d\}}{1 + \exp\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}} + \hat{\phi}\hat{v}_d\}}, \tag{13}$$

where $\hat{v}_d$ is a predictor for the random effect v_d. We describe how to calculate the corresponding predictors in Sect. 3. Finally, the plug-in predictor of Y_d is $\tilde{Y}_d^{plug} = N_d\tilde{p}_d^{plug}$.

2.3 Mean squared error estimation

In order to assess the reliability of the obtained predictions for p_d, we use the mean squared error. It is generally characterized by $MSE(\hat{p}_d) = E[(\hat{p}_d - p_d)^2]$. However, $MSE(\hat{p}_d)$ cannot be quantified directly and must be estimated instead. For this, we apply a parametric bootstrap as demonstrated by González-Manteiga et al. (2007) and Boubeta et al. (2016). It is performed as follows.

1. Fit the model to the sample and calculate the estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\phi})$.
2. Repeat B times with $b = 1, ..., B$:
   (a) Generate $v_d^{(b)} \sim N(0, 1)$, $y_d^{(b)} \sim \text{Bin}(n_d, p_d^{(b)})$, $d = 1, \ldots, D$, where $p_d^{(b)} = \frac{\exp\left\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{(b)}\right\}}{1+\exp\left\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{(b)}\right\}}$.
   (b) For each bootstrap sample, calculate the estimator $\hat{\boldsymbol{\theta}}^{(b)}$ and the EBP $\hat{p}_d^{(b)} = \hat{p}_d^{(b)}(\hat{\boldsymbol{\theta}}^{(b)})$ as stated above.
3. Output: $mse(\hat{p}_d) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{p}_d^{(b)} - p_d^{(b)}\right)^2$.

3 Penalized model parameter estimation

In this section, it is demonstrated how model parameter estimation in the area-level logit mixed model under covariate rank-deficiency is performed. The foundation of our estimation strategy is the ML-Laplace algorithm (e.g. Demidenko 2013; Hobza et al. 2018). That is to say, the integrals in the likelihood function are approximated via the Laplace method and the result is maximized with a Newton-Raphson algorithm. However, in light of the comments in Sect. 1 and prior to maximization, we extend the approximated likelihood function by the squared ℓ_2-norm of $\boldsymbol{\beta}$ to account for the negative effects of covariate rank-deficiency. With this, we obtain a penalized version approximated likelihood, which is then maximized to obtain reliable model parameter estimates. We refer to this procedure as ℓ_2-penalized approximate maximum likelihood (ℓ_2-PAML) estimation.

3.1 Laplace approximation

We first perform the Laplace approximation of the likelihood function. Let $h : R \mapsto R$ be a continuously twice differentiable function with a global maximum at x_0. This is to say, assume that $h'(x_0) = 0$ and $h''(x_0) < 0$. A Taylor series expansion of $h(x)$ around x_0 yields

$$\begin{aligned} h(x) &= h(x_0) + h'(x_0)(x - x_0) + \frac{1}{2}h''(x_0)(x - x_0)^2 + o\big(|x - x_0|^2\big) \\ &\approx h(x_0) + \frac{1}{2}h''(x_0)(x - x_0)^2. \end{aligned} \tag{14}$$

The univariate Laplace approximation is

$$\begin{aligned} \int_{-\infty}^{\infty} e^{h(x)}\,dx &\approx \int_{-\infty}^{\infty} e^{h(x_0)} \exp\Big\{ -\frac{1}{2}\big(-h''(x_0)\big)(x - x_0)^2 \Big\}\,dx \\ &= (2\pi)^{1/2}\big(-h''(x_0)\big)^{-1/2} e^{h(x_0)} \int_{-\infty}^{\infty} \frac{\exp\Big\{ -\frac{1}{2}\Big(\frac{x-x_0}{(-h''(x_0))^{-1/2}}\Big)^2 \Big\}}{(2\pi)^{1/2}\big(-h''(x_0)\big)^{-1/2}}\,dx \\ &= (2\pi)^{1/2}\big(-h''(x_0)\big)^{-1/2} e^{h(x_0)}. \end{aligned} \tag{15}$$

Let us now approximate the log-likelihood of the area-level logit mixed model. Recall that $v_1, \ldots, v_D$ are independent and identically distributed according to $v_d \sim N(0, 1)$, and that

$$y_d|_{v_d} \overset{ind}{\sim} \text{Bin}(n_d, p_d), \quad p_d = p_d(v_d) = \frac{\exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\}}{1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\}}, \quad d = 1, \ldots, D.$$

Thus, $y_1, \ldots, y_D$ are unconditionally independent with marginal probability density

$$\begin{aligned} P(y_d) &= \int_{-\infty}^{\infty} P(y_d|v_d) f(v_d)\,dv_d \\ &= \int_{-\infty}^{\infty} \left\{ \binom{n_d}{y_d} \exp\Big\{ y_d(\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\} - n_d \log\big(1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\}\big) \Big\} \right\} \\ &\quad \cdot (2\pi)^{-1/2} \exp\{-\frac{1}{2}v_d^2\}\,dv_d = (2\pi)^{-1/2}\binom{n_d}{y_d} \\ &\quad \cdot \int_{-\infty}^{\infty} \exp\Big\{ -\frac{v_d^2}{2} + y_d(\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d) - n_d \log\big(1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\}\big) \Big\}\,dv_d \\ &= (2\pi)^{-1/2}\binom{n_d}{y_d} \int_{-\infty}^{\infty} \exp\big\{h(v_d)\big\}\,dv_d, \end{aligned} \tag{16}$$

where

$$h(v_d) = -\frac{v_d^2}{2} + y_d(\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d) - n_d \log\big(1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_d\}\big). \tag{17}$$

Note that for the maximizer of $h(\cdot)$, denoted by v_{0d}, the first derivative is $h'(v_{0d}) = 0$, and the second derivative is characterized by $h''(v_{0d}) < 0$. By applying (15) in $v_d = v_{0d}$, we get

$$P(y_d) \approx \binom{n_d}{y_d} \cdot \left(1 + \phi^2 n_d p_d(v_{0d})(1 - p_d(v_{0d}))\right)^{-1/2} \cdot \exp\left\{ -\frac{v_{0d}^2}{2} + y_d(\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_{0d}) - n_d \log\left(1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_{0d}\}\right)\right\}. \quad (18)$$

From there, we can state the log-likelihood function under the model, which is given by

$$l = \sum_{d=1}^{D} l_d = \log P(y_d).$$

Using the results of the Laplace approximation, we obtain

$$l_d \approx l_{0d}(\boldsymbol{\theta}) = \log \binom{n_d}{y_d} - \frac{1}{2}\log \xi_{0d} - \frac{v_{0d}^2}{2} + y_d(\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_{0d}) n_d \log\left(1 + \exp\{\boldsymbol{x}_d\boldsymbol{\beta} + \phi v_{0d}\}\right), \quad (19)$$

where $p_{0d} = p_d(v_{0d})$ and $\xi_{0d} = 1 + \phi^2 n_d p_{0d}(1 - p_{0d})$.

3.2 ℓ_2-penalized approximate maximum likelihood

The approximated log-likelihood function is expanded by the squared ℓ_2-norm of the regression coefficients $\boldsymbol{\beta}$ to account for strong correlation between covariates in $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D$. We obtain the penalized maximum likelihood problem

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\operatorname{argmax}}\, l^{pen}(\boldsymbol{\theta}), \quad l^{pen}(\boldsymbol{\theta}) = \sum_{d=1}^{D} l_{0d}(\boldsymbol{\theta}) - \lambda \|\boldsymbol{\beta}\|_2^2, \quad (20)$$

where $l_{0d}(\boldsymbol{\theta})$ is defined in (19) and $\lambda > 0$ is a predefined tuning parameter. Maximization is performed via a Newton-Raphson algorithm. However, note that the Laplace approximations of $l_1, ..., l_D$ depends on the maximizers of $h(v_1), ..., h(v_D)$, which in turn depend on $l_1, ..., l_D$. Therefore, the maximization of (20) must contain two steps that are performed iteratively and conditional on each other. The first step is the approximation of the log-likelihood by maximizing $h(v_1), ..., h(v_D)$. The second step is the maximization of $l^{pen}(\boldsymbol{\theta})$ given the results of the first step. This is demonstrated hereafter.

Step 1: Log-likelihood approximation

In order to maximize $h(v_d)$, we need to quantify its first and second derivatives. These are

$$h'(v_d) = -v_d + \phi\{y_d - n_d p_d(v_d)\} \tag{21}$$

$$h''(v_d) = -\Big(1 + \phi^2 n_d p_d(v_d)(1 - p_d(v_d))\Big) \tag{22}$$

for all $d = 1, ..., D$. The Newton-Raphson algorithm maximizes $h(v_d) = h(v_d, \boldsymbol{\theta})$, defined in (17), for fixed $\boldsymbol{\theta} = (\boldsymbol{\beta}', \phi) = \boldsymbol{\theta}_0$. The updating equation is

$$v_d^{(k+1)} = v_d^{(k)} - \frac{h'(v_d^{(k)}, \boldsymbol{\theta}_0)}{h''(v_d^{(k)}, \boldsymbol{\theta}_0)}, \tag{23}$$

where k denotes an iteration of the procedure.

Step 2: Penalized maximization

We continue with maximizing the penalized approximate log-likelihood function. Regarding the first partial derivatives of l^{pen} with respect to $\beta_1, ..., \beta_p$ and ϕ, it holds that

$$\begin{aligned}
\frac{\partial p_{0d}}{\partial \beta_r} &= x_{dr} p_{0d}(1 - p_{0d}) = x_{dr}(p_{0d} - p_{0d}^2), \quad \frac{\partial p_{0d}}{\partial \phi} \\
&= v_{0d} p_{0d}(1 - p_{0d}) = v_{0d}(p_{0d} - p_{0d}^2), \\
\eta_{0dr} &= \frac{\partial \xi_{0d}}{\partial \beta_r} = \phi^2 n_d x_{dr}[p_{0d} - 3p_{0d}^2 + 2p_{0d}^3], \\
\eta_{0d} &= \frac{\partial \xi_{0d}}{\partial \phi} = 2\phi n_d p_{0d}(1 - p_{0d}) + \phi^2 n_d (1 - 2p_{0d}) \frac{\partial p_{0d}}{\delta \phi} \\
&= \phi n_d p_{0d}(1 - p_{0d})[2 + \phi(1 - 2p_{0d}) v_{0d}].
\end{aligned}$$

For the domain-specific likelihood component l_{0d}, this yields to

$$\frac{\partial l_{0d}}{\partial \beta_r} = -\frac{1}{2}\frac{\eta_{0dr}}{\xi_{0d}} + (y_d - n_d p_{0d}) x_{dr}, \quad \frac{\partial l_{0d}}{\partial \phi} = -\frac{1}{2}\frac{\eta_{0d}}{\xi_{0d}} + (y_d - n_d p_{0d}) v_{0d}.$$

With the application of these equations to all domain-specific likelihood components $l_{01}, ..., l_{0D}$ and the consideration of the ℓ_2-penalty, we finally obtain

$$\frac{\partial l^{pen}}{\partial \beta_r} = \sum_{d=1}^{D} \frac{\partial l_{0d}}{\partial \beta_r} - 2\lambda\beta_r, \quad \frac{\partial l^{pen}}{\partial \phi} = \sum_{d=1}^{D} \frac{\partial l_{0d}}{\partial \phi}. \tag{24}$$

For the second partial derivatives, it holds that

$$\begin{aligned}
\frac{\partial \eta_{0dr}}{\partial \beta_s} &= \phi^2 n_d x_{dr} x_{djs}[p_{0d}(1-p_{0d}) - 6p_{0d}^2(1-p_{0d}) + 6p_{0d}^3(1-p_{0d})] \\
&= \phi^2 n_d x_{dr} x_{djs} p_{0d}(1-p_{0d})[1 - 6p_{0d} + 6p_{0d}^2], \\
\frac{\partial \eta_{0dr}}{\partial \phi} &= 2\phi n_d x_{dr} p_{0d}(1-p_{0d})(1-2p_{0d}) + \phi^2 n_d x_{dr}(1 - 6p_{0d} + 6p_{0d}^2)\frac{\partial p_{0d}}{\partial \phi} \\
&= \phi n_d x_{dr} p_{0d}(1-p_{0d})[2(1-2p_{0d}) + \phi v_{0d}(1 - 6p_{0d} + 6p_{0d}^2)], \\
\frac{\partial \eta_{0d}}{\partial \beta_r} &= \phi^2 v_{0d} n_d x_{dr} p_{0d}(1-p_{0d})[1 - 6p_{0d} + 6p_{0d}^2], \\
\frac{\partial \eta_{0d}}{\partial \phi} &= 2n_d p_{0d}(1-p_{0d}) + 2\phi n_d 1 - 2p_{0d})\frac{\partial p_{0d}}{\partial \phi} \\
&\quad + 2\phi n_d(1-2p_{0d}) p_{0d}(1-p_{0d}) v_{0d} + \phi^2 n_d v_{0d}(1 - 6p_{0d} + 6p_{0d}^2)\frac{\partial p_{0d}}{\partial \phi} \\
&= n_d p_{0d}(1-p_{0d})[2 + 2\phi(1-2p_{0d})v_{0d} + 2\phi(1-2p_{0d})v_{0d} \\
&\quad + \phi^2 v_{0d}^2(1 - 6p_{0d} + 6p_{0d}^2)].
\end{aligned}$$

For the domain-specific likelihood component l_{0d}, this yields to

$$\begin{aligned}
\frac{\partial^2 l_{0d}}{\partial \beta_r^2} &= -\frac{1}{2}\frac{\frac{\partial \eta_{0dr}}{\partial \beta_r}\xi_{0d} - \eta_{0dr}^2}{\xi_d^2} - n_d x_{dr}^2 p_{0d}(1-p_{0d}), \\
\frac{\partial^2 l_{0d}}{\partial \beta_s \partial \beta_r} &= -\frac{1}{2}\frac{\frac{\partial \eta_{0dr}}{\partial \beta_s}\xi_{0d} - \eta_{0dr}\eta_{0ds}}{\xi_d^2} - n_d x_{dr} x_{djs} p_{0d}(1-p_{0d}), \\
\frac{\partial^2 l_{0d}}{\partial \phi \partial \beta_r} &= -\frac{1}{2}\frac{\frac{\partial \eta_{0dr}}{\partial \phi}\xi_{0d} - \eta_{0dr}\eta_{0d}}{\xi_d^2} - v_{0d} n_d x_{dr} p_{0d}(1-p_{0d}), \\
\frac{\partial^2 l_{0d}}{\partial \phi^2} &= -\frac{1}{2}\frac{\frac{\partial \eta_{0d}}{\partial \phi}\xi_{0d} - \eta_{0d}^2}{\xi_d^2} - v_{0d}^2 n_d p_{0d}(1-p_{0d}).
\end{aligned}$$

As for the first partial derivatives applying these equations to all domain-specific likelihood components $l_{01}, ..., l_{0D}$ and considering the ℓ_2-penalty, we end up with

$$\begin{aligned}
\frac{\partial^2 l^{pen}}{\partial \beta_r^2} &= \sum_{d=1}^{D}\frac{\partial^2 l_{0d}}{\partial \beta_r^2} - 2\lambda, \quad \frac{\partial^2 l^{pen}}{\partial \beta_s \partial \beta_r} = \sum_{d=1}^{D}\frac{\partial^2 l_{0d}}{\partial \beta_s \partial \beta_r}, \\
\frac{\partial^2 l^{pen}}{\partial \phi \partial \beta_r} &= \sum_{d=1}^{D}\frac{\partial^2 l_{0d}}{\partial \phi \partial \beta_r}, \quad \frac{\partial^2 l^{pen}}{\partial \phi^2} = \sum_{d=1}^{D}\frac{\partial^2 l_{0d}}{\partial \phi^2}.
\end{aligned} \tag{25}$$

For $r, s = 1, \ldots, p+1$, define the components of the score vector

$$U_{0r} = \frac{\partial l^{pen}}{\partial \beta_r}, \quad U_{0p+1} = \frac{\partial l^{pen}}{\partial \phi}, \tag{26}$$

as well as the Hessian matrix

$$H_{0rs} = H_{0sr} = \frac{\partial^2 l^{pen}}{\partial \beta_s \partial \beta_r}, \quad H_{rp+1} = H_{p+1r} = \frac{\partial^2 l^{pen}}{\partial \phi \partial \beta_r}, \quad H_{0p+1p+1} = \frac{\partial^2 l^{pen}}{\partial \phi^2}. \tag{27}$$

In matrix form, we have $\boldsymbol{U}_0 = \boldsymbol{U}_0(\boldsymbol{\theta}) = \underset{1 \le r \le p+1}{\text{col}}(U_{0r})$ and $\boldsymbol{H}_0 = \boldsymbol{H}_0(\boldsymbol{\theta}) = (H_{0rs})_{r,s=1,\ldots,p+1}$. The Newton-Raphson algorithm maximizes $l^{pen}(\boldsymbol{\theta})$, with fixed $v_d = v_{0d}$, $d = 1, \ldots, D$. Let k denote the index of iterations. The corresponding updating equation is

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \boldsymbol{H}_0^{-1}(\boldsymbol{\theta}^{(k)}) \boldsymbol{U}_0(\boldsymbol{\theta}^{(k)}). \tag{28}$$

Complete ℓ_2-PAML algorithm

The final algorithm containing both steps is performed as follows.

1. Set the initial values $k = 0$, $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(-1)} = \boldsymbol{\theta}^{(0)} + \mathbf{1}$, $v_d^{(0)} = 0$, $v_d^{(-1)} = 1$, $d = 1, \ldots, D$.
2. Until $\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}\|_2 < \varepsilon_1$, $|v_d^{(k)} - v_d^{(k-1)}| < \varepsilon_2$, $d = 1, \ldots, D$, do
   (a) Apply algorithm (23) with seeds $v_d^{(k)}$, $d = 1, \ldots, D$, convergence tolerance ε_2 and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ fixed. Output: $v_d^{(k+1)}$, $d = 1, \ldots, D$.
   (b) Apply algorithm (28) with seed $\boldsymbol{\theta}^{(k)}$, convergence tolerance ε_1 and $v_{0d} = v_d^{(k+1)}$ fixed, $d = 1, \ldots, D$. Output: $\boldsymbol{\theta}^{(k+1)}$.
   (c) $k \leftarrow k + 1$.
3. Output: $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(k)}$, $\hat{v}_d = v_d^{(k)}$, $d = 1, \ldots, D$.

We remark that the output of the ℓ_2-PAML algorithm gives estimates $\hat{\boldsymbol{\theta}}$ of the model parameters $\boldsymbol{\theta}$ and mode predictions $\hat{v}$ of the random effects v_d, $d = 1, ..., D$.

3.3 Tuning parameter choice and information criterion

In the technical descriptions of Sect. 3.2, we assumed that the tuning parameter λ had been defined prior to model parameter estimation. In practice, it has to be found empirically from the sample data. Note that this aspect is crucial for the effectiveness of the proposed method. On the one hand, if λ is chosen too small, the ℓ_2-PAML approach cannot sufficiently stabilize model parameter estimates in the presence of covariate rank-deficiency. On the other hand, if λ is chosen too large, the shrinkage induced by penalization dominates the optimization problem and resulting model parameter estimates are heavily biased. Finding an appropriate value for the tuning parameter is often done via grid search, as can be seen for instance in Bergstra and Bengio (2012) and Chicco (2017). We define a sequence of candidate values $\{\lambda_q\}_{q=1}^{Q}$, where $\lambda_q > \lambda_{q+1}$. For each candidate value λ_q model parameter estimation as demonstrated in Sect. 3.2 is performed. The results of model parameter estimation have to be evaluated by a

suitable goodness-of-fit measure. For our application, we choose the non-corrected Bayesian information criterion (BIC; Schwarz 1978). Alternative measures would be the generalized cross-validation criterion (Craven and Wahba 1979) or the Akaike information criterion (Akaike 1974). For given candidate value λ_q, let $\hat{\boldsymbol{\beta}}(\lambda_q)$ and $\hat{\phi}(\lambda_q)$ be the estimators of $\boldsymbol{\beta}$ and ϕ, respectively. Further, let $\hat{v}_d(\lambda_q)$ be the mode predictor of v_d. The Laplace non-corrected BIC is given by

$$BIC(\lambda_q) = p\log(D) - 2l^{app}(\lambda_q), \tag{29}$$

where the second term is the Laplace approximation (19) to the log-likelihood, that is

$$\begin{aligned} l^{app}(\lambda_q) &= l^{app}\left(\hat{\boldsymbol{\beta}}(\lambda_q), \hat{\phi}(\lambda_q), \hat{v}_1(\lambda_q), \ldots, \hat{v}_D(\lambda_q)\right) \\ &= \sum_{d=1}^{D} \log\binom{n_d}{y_d} + \sum_{d=1}^{D}\Bigg\{ -\frac{1}{2}\log\hat{\xi}_d(\lambda_q) - \frac{\hat{v}_d(\lambda_q)^2}{2} \\ &\quad + \Big\{ y_d(\boldsymbol{x}_d\hat{\boldsymbol{\beta}}(\lambda_q) + \hat{\phi}(\lambda_q)\hat{v}_d(\lambda_q)) \\ &\quad - n_d\log\left(1 + \exp\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}(\lambda_q) + \hat{\phi}(\lambda_q)\hat{v}_d(\lambda_q)\}\right)\Big\}\Bigg\}, \end{aligned}$$

where

$$\begin{aligned} \hat{\xi}_d(\lambda_q) &= 1 + \hat{\phi}(\lambda_q)^2 n_d \hat{p}_d(\lambda_q)(1 - \hat{p}_d(\lambda_q)), \\ \hat{p}_d(\lambda_q) &= \frac{\exp\left\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}(\lambda_q) + \hat{\phi}(\lambda_q)\hat{v}_d(\lambda_q)\right\}}{1 + \exp\left\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}(\lambda_q) + \hat{\phi}(\lambda_q)\hat{v}_d(\lambda_q)\right\}}. \end{aligned}$$

For all $\lambda_1, ..., \lambda_Q$, the following algorithm is performed:

1. Apply the ℓ_2-PAML algorithm to obtain $\hat{\boldsymbol{\theta}}(\lambda_q)$ and $\hat{v}_1(\lambda_q), ..., \hat{v}_D(\lambda_q)$.
2. Calculate $\hat{p}_d(\lambda_q)$ and $\hat{\xi}_d(\lambda_q)$, $d = 1, ..., D$.
3. Calculate $BIC(\lambda_q)$ according to (29).

After the algorithm is finalized, the optimal tuning parameter λ^{opt} can be defined as the candidate value that minimizes the BIC.

However, due to the non-convexity of the underlying optimization problem for ℓ_2-PAML estimation, the behavior of the BIC along the tuning parameter sequence can be volatile to the extent that it may be characterized by multiple local minima. Therefore, we further apply cubic spline smoothing by defining $BIC(\lambda) = f(\lambda) + \epsilon_q$, where $f(\lambda)$ is a twice differentiable function and $\epsilon_q \sim N(0, \psi)$. The cubic spline estimate $\hat{f}$ of the function f is obtained from solving the optimization problem

$$\min_{f\in\mathcal{F}} \sum_{q=1}^{Q} \left[BIC(\lambda_q) - f(\lambda_q)\right]^2 + \delta \int f''(\lambda)^2 \, \mathrm{d}\lambda, \tag{30}$$

where $\mathcal{F} = \{f : f \text{ is twice differentiable}\}$ denotes the class of twice differentiable functions and $\delta > 0$ is a smoothing parameter. After $\hat{f}$ has been obtained, the optimal tuning parameter value λ^{opt} is defined as the minimizer of the smoothed function, that is

$$\lambda^{opt} = \underset{\lambda \in \{\lambda_q\}_{q=1}^{Q}}{\operatorname{argmin}} \ \hat{f}(\lambda). \tag{31}$$

4 Simulation

4.1 Setup

Hereafter, the performance of the ℓ_2-PAML approach is evaluated under controlled conditions. For this, we conduct a Monte Carlo simulation study with $K = 500$ iterations that are indexed by $k = 1, ..., K$. We generate synthetic data according to

$$y_d \sim \text{Bin}(n_d, p_d), \quad p_d = \frac{\exp\left\{\beta_0 + \boldsymbol{x}_d \boldsymbol{\beta}_1 + \phi v_d\right\}}{1 + \exp\left\{\beta_0 + \boldsymbol{x}_d \boldsymbol{\beta}_1 + \phi v_d\right\}},$$
$$\beta_0 = -0.2, \ \boldsymbol{\beta}_1 = 0.3\mathbf{1}_5, \quad d = 1, ..., D,$$

where $n_d = 100$, $\mathbf{1}_5$ as column vector of five ones, and $\phi = 0.4$. The random effect v_d is drawn from a standard normal, as defined in Sect. 2.1. For the covariate vector $\boldsymbol{x}_d$, we consider four different settings {A, B, C, D} with respect to the dependency between the auxiliary variables. This is done in order to test the methodology under different covariate correlation situations. In the A-setting, we have orthogonal covariates that are generated according to $x_{rd} \sim U(0.7, 1.2)$, $r = 1, ..., 5$. For the remaining three settings, we choose

$$x_{1d} \sim U(0.7, 1.2), \quad x_{rd} = \alpha(z_d + \rho x_{1d}), \quad z_d \sim U(0, 0.2), \quad r = 2, ..., 5,$$

where ρ is a parameter controlling the dependency between x_{1d} and x_{rd}, and α is a parameter harmonizing the variance of the random variables over settings. In the B-setting, there is medium correlation with 20-50% on a percentage scale for the product-moment correlation coefficient. For the C-setting, we have correlation with about 50-75%. And in the D-setting, we have a strong correlation with 80-90%. Note that the latter mimics situations of quasi rank-deficiency, which are of special interest. In addition to covariate correlation, we let the total number of areas D vary over scenarios in order to evaluate the method under different degrees of freedom. Overall, we consider 8 simulation scenarios:

A.1 : $D = 50$,	**A.2** : $D = 100$,
B.1 : $D = 50, \rho = 0.3, \alpha = 2.0$,	**B.2** : $D = 100, \rho = 0.3, \alpha = 2.0$,
C.1 : $D = 50, \rho = 0.9, \alpha = 1.5$,	**C.2** : $D = 100, \rho = 0.9, \alpha = 1.5$,
D.1 : $D = 50, \rho = 1.5, \alpha = 0.7$,	**D.2** : $D = 100, \rho = 1.5, \alpha = 0.7$.

The objective is to estimate the domain proportion p_d, $d = 1, ..., D$. We compare two model parameter estimation approaches for the logit mixed model described in Sect. 2.1: a non-penalized approach that is obtained from maximizing l^{app} (Laplace-ML), and the ℓ_2-penalized approach through maximizing l^{pen} (ℓ_2-PAML), as described in Sect. 3. We evaluate the simulation outcomes with respect to three aspects: (i) model parameter estimation, (ii) domain proportion prediction, and (iii) MSE estimation based on the parametric bootstrap in Sect. 2.3. The results are summarized in the following subsections.

4.2 Model parameter estimation results

The target of this subsection is to study the fitting behavior of the ℓ_2-PAML algorithm. Define $\boldsymbol{\theta} := (\beta_0, \boldsymbol{\beta}_1', \phi)$. For a given estimator $\hat{\theta}_r \in \hat{\boldsymbol{\theta}}$ of the model parameter θ_r, $r = 1, ..., p + 1$, we consider the following performance measures:

$$Bias(\hat{\theta}_r) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_r^{(k)} - \theta_r\right), MSE(\hat{\theta}_r) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_r^{(k)} - \theta_r\right)^2, \tag{32}$$

where $\theta_r^{(k)}$ is the value that $\hat{\theta}_r$ takes in the k-th iteration of the simulation and θ_r denotes the true value. As $\theta_r = 0.3$ for all components of $\boldsymbol{\beta}_1$, we average the performance measures for the regression parameters. Table 1 contains the results for model parameter estimation.

We start with the regression parameters $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')'$. It can be seen that the ℓ_2-PAML algorithm obtains more efficient estimates than the ML-Laplace approach. Its MSE is significantly smaller in all considered scenarios. The largest efficiency gains are obtained in the D-scenarios, which include strong covariate correlation. This could be expected from theory, as the ℓ_2-penalty was introduced by Hoerl and Kennard (1970) in order to improve the fitting behavior in these settings. However, we also see that under orthogonal covariates (A-scenarios), the ℓ_2-PAML algorithm still outperforms the ML-Laplace approach. This is because approximate likelihood inference introduces additional uncertainty to model parameter estimation. Here, the ℓ_2-penalty stabilizes the shape of the objective function, which allows for efficiency gains even without covariate correlation. Yet, the increased efficiency comes at the cost of an increased bias. The slope parameters $\boldsymbol{\beta}_1$, which are penalized while applying the ℓ_2-PAML algorithm, are estimated with larger bias relative to the ML-Laplace method. Please note that this is in line with theory. Hoerl and Kennard (1970) showed that the ℓ_2-penalty affects the bias-variance trade-off the researcher typically encounters in ML estimation. It increases the bias in order to reduce the variance, which ultimately allows for a smaller MSE when the regularization parameter λ is chosen appropriately.

This is also becomes evident when looking at the distribution of regression parameter estimates. Figure 1 shows boxplots of the absolute deviation $|\hat{\beta}_r - \beta_r|$, $\beta_r \in \boldsymbol{\beta}_1$, over all Monte Carlo iterations and for different simulation scenarios. In each quarter, the distribution yielded by the ℓ_2-PAML algorithm is displayed on the left, while the the one obtained by the ML-Laplace algorithm is located on the right. We see that the

Table 1 Model Parameter Estimation Results

Scen	Method	$Bias(\beta_0)$	$MSE(\beta_0)$	$Bias(\boldsymbol{\beta}_1)$	$MSE(\boldsymbol{\beta}_1)$	$Bias(\phi)$	$MSE(\phi)$
A.1	ML-Laplace	−0.2760	2.3958	0.0059	0.4972	0.1691	0.0344
A.1	ℓ_2-PAML	0.6752	1.4878	−0.1889	0.1908	0.1832	0.0416
A.2	ML-Laplace	−0.3861	1.1191	0.0317	0.2364	0.1870	0.0381
A.2	ℓ_2-PAML	0.5021	0.9434	−0.1522	0.1215	0.1985	0.0430
B.1	ML-Laplace	−0.3128	0.6994	0.0132	0.9795	0.1699	0.0345
B.1	ℓ_2-PAML	0.2271	0.5077	−0.1197	0.3004	0.1819	0.0406
B.2	ML-Laplace	−0.3086	0.3347	0.0136	0.4284	0.1898	0.0390
B.2	ℓ_2-PAML	0.1501	0.2963	−0.1007	0.1534	0.2014	0.0441
C.1	ML-Laplace	−0.3380	0.8412	0.0060	3.1709	0.1621	0.0325
C.1	ℓ_2-PAML	0.4065	0.8252	−0.0979	0.7336	0.1806	0.0409
C.2	ML-Laplace	−0.3096	0.3974	0.0087	1.7727	0.1813	0.0359
C.2	ℓ_2-PAML	0.2329	0.3875	−0.0674	0.3062	0.1931	0.0413
D.1	ML-Laplace	−0.3345	0.7745	0.0186	11.7489	0.1641	0.0328
D.1	ℓ_2-PAML	0.2697	0.6943	−0.0939	3.1089	0.1852	0.0416
D.2	ML-Laplace	−0.2907	0.3461	0.0128	5.6289	0.1868	0.0380
D.2	ℓ_2-PAML	0.1016	0.3210	−0.0615	1.7118	0.2021	0.0448

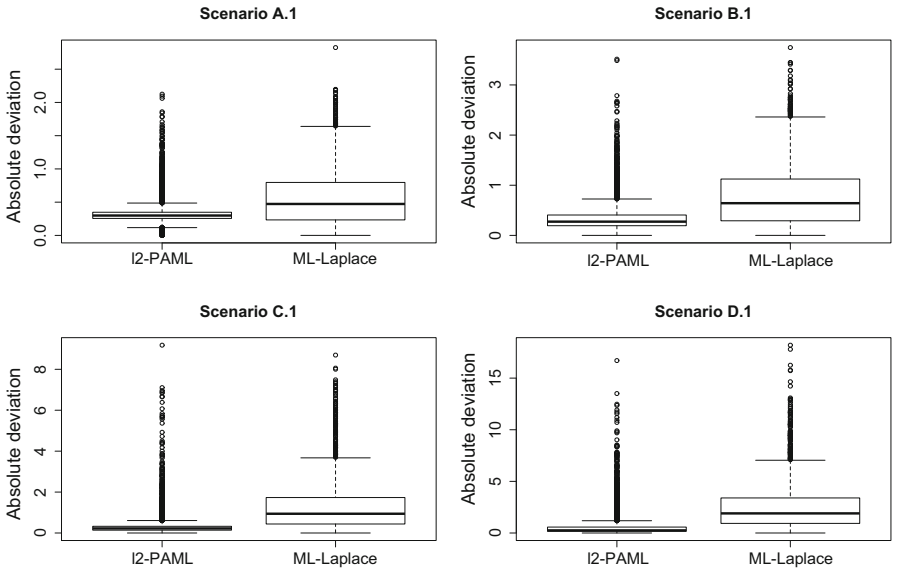

Fig. 1 Absolute deviation of regression parameter estimates

boxes and whiskers of the ℓ_2-PAML algorithm are much shorter than those of the ML-Laplace method. This implies that the deviations from the true value are much smaller under penalization for the vast majority of cases. Accordingly, the fitting behavior is overall stabilized.

Concerning ϕ, the results are different. The standard deviation parameter estimation is not influenced by the covariate correlation. An intuitive explanation for this phenomenon is that p_{0d} is not affected by the collinearity of $\boldsymbol{x}_d$, and that the diagonal element $H_{0p+1p+1}$ of the Hessian matrix depends on $\boldsymbol{x}_d$ only through p_{0d}. This is why, we expect that the asymptotic behavior of the ML-Laplace and ℓ_2-PAML estimators of ϕ will be not (or almost not) affected by the covariate correlation.

Concerning the comparison of the two fitting algorithms, the ℓ_2-PAML approach increases the efficiency of regression parameter estimation. On the other hand, the efficiency of standard deviation parameter estimation is impaired relative to the ML-Laplace approach. In general, both methods overestimate the true value. This is likely due the involved Laplace approximation in both algorithms. It is known to induce bias to model parameter estimation, as for instance addressed by Jiang (2007), p. 131. However, the bias for the ℓ_2-PAML algorithm is larger, as it implements additional shrinkage of the regression parameters through the ℓ_2-penalty. The regression parameter estimates are drawn to zero (to some extent), which causes a larger proportion of the target variable's variance to be attributed to the random effect. This leads to a stronger overestimation of the random effect standard deviation. Nevertheless, we will see in the next subsection that the efficiency advantage in regression parameter estimation overcompensates the loss in standard devation parameter estimation accuracy.

4.3 Domain proportion prediction results

The target of this subsection is to investigate the behavior of the EBP of p_d, $d = 1, ..., D$. We consider absolute bias, MSE, relative absolute bias, and relative root mean squared error as performance measures. For a domain proportion prediction in the k-th iteration of the simulation study, define

$$\bar{p}_d = \frac{1}{K}\sum_{k=1}^{K} p_d^{(k)},\ RB_d = \frac{\sum_{k=1}^{K} |\hat{p}_d^{(k)} - p_d^{(k)}|}{K|\bar{p}_d|},$$
$$RE_d = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{p}_d^{(k)} - p_d^{(k)})^2}}{|\bar{p}_d|},\ d = 1, \ldots, D.$$

Further, let

$$B_d = \frac{1}{K}\sum_{k=1}^{K} |\hat{p}_d^{(k)} - p_d^{(k)}|,\ E_d = \frac{1}{K}\sum_{k=1}^{K}(\hat{p}_d^{(k)} - p_d^{(k)})^2,\ d = 1, \ldots, D.$$

The performance measures are then given by

$$ABias = \frac{1}{D}\sum_{d=1}^{D} B_d,\ RABias = \frac{1}{D}\sum_{d=1}^{D} RB_d,$$
$$MSE = \frac{1}{D}\sum_{d=1}^{D} E_d,\ RRMSE = \frac{1}{D}\sum_{d=1}^{D} RE_d.$$

The results obtained from the simulation study are summarized in Table 2. We observe that ℓ_2-PAML improves domain total prediction performance in terms of all considered performance measures and for all implemented simulation scenarios, including those without covariate correlation. This is in line with the simulation results for model parameter estimation from the last subsection. The ℓ_2-penalty stabilizes the estimation performance even for orthogonal covariates due to the necessary Laplace approximation. However, the strongest efficiency gains in terms of the MSE relative to the ML-Laplace algorithm are obtained in the C- and D-scenarios, where we have strong covariate correlation. Against the backhground of Hoerl and Kennard (1970), this could be expected from theory, as ℓ_2-penalization is known to be particularly useful in the presence of (quasi-)multicollinearity. Overall, we can conclude that the ℓ_2-PAML algorithm not only improves model parameter estimation, but also domain total prediction in any setting.

4.4 Mean squared error estimation results

The target of this subsection is to study the performance of the parametric bootstrap for MSE estimation. We employ $B = 500$ bootstrap replicates in order to approximate

Table 2 Domain Proportion Prediction Results

Scen	Method	ABias	MSE	RABias	RRMSE
A.1	ML-Laplace	0.033342	0.001781	0.047429	0.059960
A.1	ℓ_2-PAML	0.032920	0.001735	0.046830	0.059188
A.2	ML-Laplace	0.033094	0.001749	0.047128	0.059465
A.2	ℓ_2-PAML	0.032911	0.001729	0.046870	0.059132
B.1	ML-Laplace	0.034811	0.001922	0.053363	0.066946
B.1	ℓ_2-PAML	0.034416	0.001879	0.052758	0.066201
B.2	ML-Laplace	0.034372	0.001872	0.052849	0.066259
B.2	ℓ_2-PAML	0.034180	0.001851	0.052557	0.065877
C.1	ML-Laplace	0.028710	0.001344	0.036400	0.046199
C.1	ℓ_2-PAML	0.028228	0.001300	0.035794	0.045434
C.2	ML-Laplace	0.028748	0.001355	0.036663	0.046597
C.2	ℓ_2-PAML	0.028539	0.001333	0.036399	0.046227
D.1	ML-Laplace	0.032102	0.001652	0.044746	0.056371
D.1	ℓ_2-PAML	0.031651	0.001605	0.044116	0.055569
D.2	ML-Laplace	0.032197	0.001663	0.045118	0.056794
D.2	ℓ_2-PAML	0.031972	0.001640	0.044802	0.056401

the prediction uncertainty under the model. For a MSE estimate in the k-th iteration of the simulation study, define

$$MSE_d = \frac{1}{K}\sum_{k=1}^{K}(\hat{Y}_d^{(k)} - Y_d^{(k)})^2, \quad mse_d = \frac{1}{K}\sum_{k=1}^{K} mse(\hat{Y}_d^{(k)}), \quad mse = \frac{1}{D}\sum_{d=1}^{D} mse_d,$$

where $\hat{Y}_d^{(k)}$ and $mse(\hat{Y}_d^{(k)})$ are the EBP of Y_d and its bootstrap MSE estimator (see Sect. 2.3), respectively. We consider the following performance measures

$$RBias = \frac{1}{D}\sum_{d=1}^{D}\frac{mse_d - MSE_d}{MSE_d}, \quad RRMSE = \frac{1}{D}\sum_{d=1}\frac{\sqrt{\frac{1}{D}\sum_{d=1}^{D}(mse_d - MSE_d)^2}}{MSE_d}.$$

Table 3 summarizes the simulation results. We see that the parametric bootstrap estimator shows a decent performance overall. There is a slight tendency for underestimation. However, with a relative bias of less then 4.3% for approximate likelihood inference with a generalized linear mixed model, this is negligible. With regards to the RRMSE, we see that the parametric bootstrap is more efficient under orthogonal covariates (A-scenarios) and medium correlation (B-scenarios). In the C- and D-scenarios, which employ stronger covariate correlation, the RRMSE becomes larger. This is in line with the results of Sect. 4.2. In these scenarios, the model parameter estimates are subject to larger variation, which affects the bootstrap due to its parametric construction. Yet,

Table 3 Mean Squared Error Estimation Results

Scen	MSE	mse	RBias	RRMSE
A.1	0.001735	0.001687	−0.027801	0.097148
A.2	0.001729	0.001676	−0.030432	0.089698
B.1	0.001879	0.001851	−0.014858	0.090956
B.2	0.001851	0.001842	−0.004771	0.085181
C.1	0.001300	0.001275	−0.019039	0.171667
C.2	0.001333	0.001276	−0.042769	0.168694
D.1	0.001605	0.001613	0.004951	0.124678
D.2	0.001640	0.001610	−0.018371	0.119636

with respect to practice, a RRMSE ranging from 8.5% to 17.2% is a solid result for uncertainty estimation.

5 Application

5.1 Data description and model specification

In what follows, we apply the ℓ_2-PAML approach to estimate the regional prevalence of multiple sclerosis in Germany. For this, we consider the German population of the year 2017. It is segmented into 401 administrative districts and contains about 82 million individuals. The districts correspond to the domains in accordance with Sect. 2.1. The required demographic information is retrieved from the German Federal Statistical Office and based on the methodological standards described in Statistisches Bundesamt (2016). As model response y, we define a binary variable with realizations

$$y_{id} = \begin{cases} 1 & \text{person has multiple sclerosis} \\ 0 & \text{else} \end{cases}$$

for some $i \in U_d$. The objective is to estimate $p_d = Y_d/N_d$ with $Y_d = \sum_{i \in U_d} y_{id}$ for all German districts. In order to define whether a person has multiple sclerosis, we rely on an intersectoral disease profile provided by the Scientific Institute Institute of the AOK (WIdO). It is based on multiple aspects, including medical descriptions, inpatient diagnoses, and ambulatory diagnoses. The necessary sample counts for y are based on health insurance records provided by the AOK. In particular, we use district-level prevalence figures of the AOK insurance population in 2017 that are based on the intersectoral disease profiles. The AOK insurance population is the biggest statutory health insurance population of the country with roughly 26 million individuals in 2017 (AOK Bundesverband 2018). Note that the German health insurance system has a rather unique separation between statutory and private health insurance. Usually, this has to be accounted for in order to produce reliable prevalence estimates. However, Burgard et al. (2019) showed that model-based inference using covariate data with sufficient explanatory power can overcome this issue.

As auxiliary data source, we use district-level inpatient diagnosis frequencies of the German DRG-Statistics that are provided by the German Federal Statistical Office (Statistisches Bundesamt 2017). The data set contains figures on how often a given disease has been recorded in hospitals within a year. Both main and secondary diagnoses are considered. With respect to diagnosis grouping, the records are provided on the ICD-3 level (World Health Organization 2018). Note that the DRG-Statistics are a full census of all German hospitals. Thus, the corresponding records cover the entire population, as required for the model derivation in Sect. 2.1. However, a drawback of the data set's richness is that we have to choose a suitable set of predictors x out of approximately 3 000 potential covariates. Naturally, it is not feasible to apply an exhaustive stepwise strategy that is often used in the context of variable selection, as for instance demonstrated by Yamashita et al. (2007).

Instead, we apply a heuristic strategy based on the premise that the objective is to find a covariate subset with sufficient explanatory power for our purpose. First, we isolate the 20 variables of the DRG-Statistics that have the strongest correlation with the AOK records on *G35*, which is multiple sclerosis on the ICD-3 level. The variables are arranged in decreasing order with respect to their correlation. Next, we use the ℓ_2-PAML algorithm from Sect. 3.2 to perform model parameter estimation for p covariates, where $p \in \{2, 3, ..., 20\}$. That is to say, we start with the 2 covariates that have the strongest correlation to *G35*, and then sequentially increase the number of predictors up to 20. For every result of model parameter estimation, we calculate the Laplace non-corrected BIC in (29). Then, we select the covariate subset that corresponds to the model fit which minimizes the BIC. The BIC curve over all considered covariate set cardinalities is displayed in Fig. 2. We see that the curve has an odd evolution over the covariate sets. This can be attributed to three reasons. Firstly, due to the non-linearity of the link function, the covariate sorting is guaranteed to organize the covariates in descending order with regards to their relevance for the target variable. Secondly, due to the strong correlation between them, the covariate contributions interfer with each other. That is to say, when including an additional covariate into the active set, the contributions of the previously contained covariates can change considerably. And finally, as already addressed in Sect. 3.3, the non-convexity of the optimization problem further leads to irregularities in the BIC curve.

Despite these issues, the BIC curve has a clear minimum that is located at $p = 9$. Therefore, we isolate the 9 DRG-Statistics variables that have the strongest correlation with the AOK records on *G35*. Thereafter, we perform a parametric bootstrap to estimate the standard deviation of each model parameter estimate $\hat{\theta}_j \in \hat{\boldsymbol{\theta}}$, $j = 1, ..., p+1$, to evaluate its significance in terms of the p-value. The parametric bootstrap is described as follows:

1. Fit the model to the sample and calculate the estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\phi})$.
2. Repeat B times with $b = 1, ..., B$:
   (a) Generate $v_d^{(b)} \sim N(0, 1)$, $y_d^{(b)} \sim \text{Bin}(n_d, p_d^{(b)})$, $d = 1, \ldots, D$, where $p_d^{(b)} = \frac{\exp\left\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}+\hat{\phi}v_d^{(b)}\right\}}{1+\exp\left\{\boldsymbol{x}_d\hat{\boldsymbol{\beta}}+\hat{\phi}v_d^{(b)}\right\}}$.
   (b) For each bootstrap sample, calculate the estimator $\hat{\boldsymbol{\theta}}^{(b)}$.

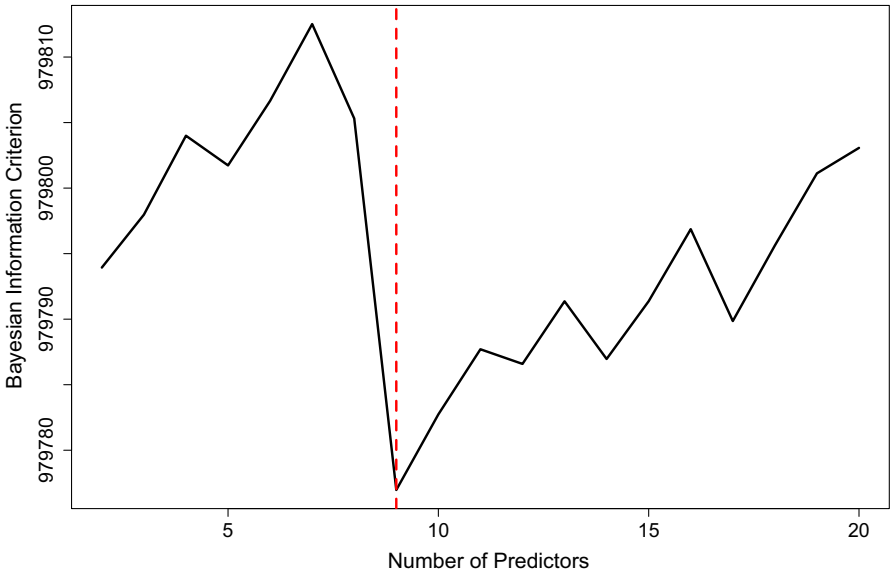

Fig. 2 BIC over covariate set cardinalities

3. Output: $sd(\hat{\theta}_j) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_j^{(b)} - \frac{1}{B}\sum_{k=1}^{B}\hat{\theta}_j^{(k)}\right)^2}$, $j = 1, ..., p+1$.

Based on the estimated standard deviations, we calculate test statistics for a sequence of t-tests under the null hypothesis $H_0 : \theta_j = 0$, $j = 1, ..., p+1$. For a given $\theta_j \in \boldsymbol{\theta}$, the test statistic is given by $t_j = \hat{\theta}_j / sd(\hat{\theta}_j)$ and follows a standard normal distribution. The test statistic values are located in the pdf of the standard normal to obtain their respective p-values. We delete every predictor that corresponds to a model parameter that is not relevant on at least a 10% significance level. The entire procedure is summarized hereafter:

1. Find the 20 covariates with the strongest correlation to y
2. Perform model parameter estimation for $p \in \{2, 3, ..., 20\}$ predictors
3. Find the number of predictors that minimizes the BIC
4. For the BIC-minimal predictor set, perform a parametric bootstrap to estimate standard deviations for the model parameter estimates
5. Perform t-tests to evaluate their significance and delete insignificant predictors

The proposed strategy yields us the final covariate set x which consists of $p = 5$ predictors. The selected covariates are briefly characterized as follows:

- X_1: G43 (Migraine, secondary diagnosis)
- X_2: M20 (Acquired deformities of fingers and toes, main diagnosis)
- X_3: E66 (Overweight and obesity, main diagnosis)
- X_4: E04 (Other nontoxic goiter, main diagnosis)
- X_5: G35 (Multiple sclerosis, secondary diagnosis)

Please note that the association of these variables with multiple sclerosis is the result of district-level correlation. It does not directly imply person-level comorbidities in

Table 4 Estimation results for final model specification

Parameter	Estimate	Std.Dev.	p-value	95%-Conf.Int.
β_0	−5.79912	0.00640	0.00000	[−5.81167; −5.78657]
β_1	−0.00092	0.00009	0.00000	[−0.00110; −0.00073]
β_2	0.00026	0.00009	0.00380	[0.00009; 0.00043]
β_3	−0.00030	0.00007	0.00001	[−0.00043; −0.00017]
β_4	−0.00081	0.00006	0.00000	[−0.00093; −0.00068]
β_5	0.00154	0.00009	0.00000	[0.00136; 0.00173]
ϕ	0.10939	0.00551	0.00000	[0.09860; 0.12019]

a medical sense. Applying the ℓ_2-PAML algorithm on the final covariate set yields us the final model specification that we use for regional prevalence estimation. It is summarized in Table 4. The confidence intervals for the parameters are calculated according to $\hat{\theta}_j \pm t_{(D,1-\alpha/2)} sd(\hat{\theta}_j)$, $j = 1, ..., p+1$, where $t_{(\cdot)}$ is the corresponding quantile of t-distribution with D degrees of freedom and significance level α. The BIC value of the upper model specification is 979754 and therefore even better than the optimal fit with $p = 9$ in Fig. 2. This suggests that the used model specification was a reasonable choice given the considered data basis. Further, observe that the estimated value for the standard deviation parameter ϕ is considerably larger than all slope parameters $\beta_1, ..., \beta_5$. This implies that the random effects $v_1, ..., v_D$ are clearly evident in the empirical distribution of $p_1, ..., p_D$. Therefore, we can conclude that using a mixed effect model in this context was a necessary choice.

We further look at the internal correlation structure of the considered predictors in order to assess the demand for ℓ_2-penalization in the application. For this, we look at the empirical correlation matrix for the five selected DRG-Statistics variables. It is given as follows:

$$\varrho_{xx} = \begin{pmatrix} 1.00 & 0.95 & 0.93 & 0.85 & 0.94 \\ 0.95 & 1.00 & 0.94 & 0.87 & 0.95 \\ 0.93 & 0.94 & 1.00 & 0.88 & 0.95 \\ 0.85 & 0.87 & 0.88 & 1.00 & 0.88 \\ 0.94 & 0.95 & 0.95 & 0.88 & 1.00 \end{pmatrix}$$

We observe that (beside the main diagonal elements), the correlation values range from 0.85 to 0.95, or 85% to 95% on a percentage scale. This suggests that the internal correlation structure is very strong and comparable to the D-scenarios of our simulation study. Therefore, we conclude that using ℓ_2-penalization is reasonable in this context. However, note that some of this correlation is due to the size as a result of district-level aggregation. Again, this does not directly resemble medical comorbidity on an individual level.

Table 5 Quantiles of the EBP distributions

Method	Min	0.25	0.50	Mean	0.75	Max
ML-Laplace	0.171%	0.259%	0.297%	0.301%	0.336%	0.513%
ℓ_2-PAML	0.209%	0.269%	0.295%	0.300%	0.324%	0.471%

5.2 Results

Let us now investigate the results of prevalence estimation. The national prevalence $\sum_{d=1}^{D} Y_d / \sum_{d=1}^{D} N_d \cdot 100\%$ is estimated at 0.296%. Based on the parametric bootstrap, we calculate a 95% confidence interval of [0.293%; 0.300%]. This implies that the estimated total number of persons with multiple sclerosis ranges approximately from 239 000 to 246 000, which is in line with reference figures on this topic. The Central Research Institute of Ambulatory Health Care in Germany estimated that in 2017 about 240 000 individuals had multiple sclerosis (Müller 2018). The regional distribution of prevalence estimates on the district-level is displayed in Fig. 3. We observe a prevalence discrepancy between the western and eastern parts of Germany. The estimated prevalence in western Germany are overall higher than in eastern Germany. Further, we observe regional clustering with higher prevalence in the central-northern and central-southern parts of Germany. This is also consistent with reference studies. Similar patterns have been found by Central Research Institute of Ambulatory Health Care in Germany (Müller 2018) and Petersen et al. (2014). Overall, the estimates are plausible in both level and geographic distribution.

Figure 3 shows the distributions of district-level prevalence estimates for the EBPs under both ℓ_2-PAML and the classical ML-Laplace method. The ML-Laplace results are displayed in black, the ℓ_2-PAML results are plotted in red. We see that the means of the distributions are almost identical. However, the ℓ_2-PAML distribution shows considerably less variance than the ML-Laplace distribution. This is in line with both theory and the simulation study, which both suggest stabilizing effects through ℓ_2-penalization.

This is further evident when looking at the summarizing quantiles of both predictive distributions. They are displayed in Table 5 . We see that the ℓ_2-PAML estimates are more more focussed around the mean and do not show as strong of outliers at the tails of the distribution compared to ML-Laplace.

Figure 5 displays the root mean squared error estimates $rmse(\hat{p}_d) = \sqrt{mse(\hat{p}_d)}$ for the prevalence estimates in Fig. 3, where $mse(\hat{p}_d)$ is obtained from the parametric bootstrap procedure described in Sect. 2.3. It becomes evident that there are no obvious spatial patterns in the RMSE estimates. We neither observe a particular dependency on the domain sizes nor on the prevalence estimates themselves. However, with respect to the overall level of RMSE estimates, we can conclude that our estimates are more efficient than direct estimates $\hat{p}_d^{dir} = y_d/n_d$, $d = 1, ..., D$, that are exclusively obtained from the health insurance records. Their standard deviation is given by $sd(\hat{p}_d^{dir}) = \sqrt{\hat{p}_d^{dir}(1 - \hat{p}_d^{dir})}$.

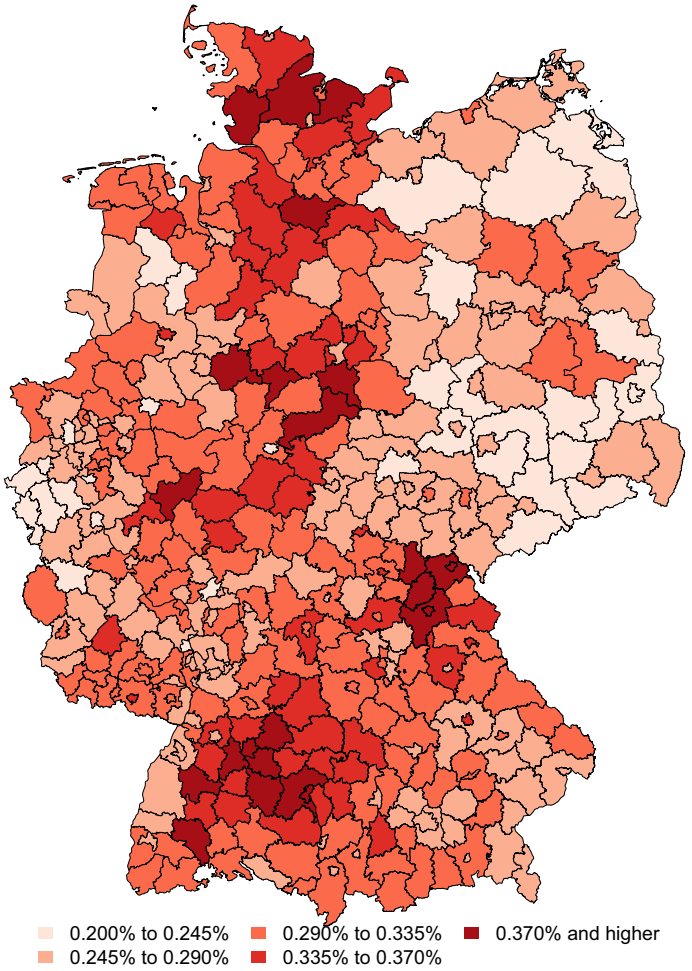

Fig. 3 Results of prevalence estimation

A one-to-one comparison of $rmse(\hat{p}_d)$ and $sd(\hat{p}_d^{dir})$ per domain is visualized in Fig. 6. The ordinate measures $sd(\hat{p}_d^{dir})$ and the abscissa measures $rmse(\hat{p}_d)$. The red line marks the bisector, which indicates equality between the two. We observe that $rmse(\hat{p}_d)$ is always smaller than $sd(\hat{p}_d^{dir})$ by quite a margin. Thus, given the reasonable performance of the parametric bootstrap for MSE estimation in the simulation, we can conclude that our estimates mark an improvement over the direct estimates. There is a slight positive relation between the two measures. That is to say, a relatively large $sd(\hat{p}_d^{dir})$ is accompanied by a relatively large $rmse(\hat{p}_d)$ on expectation. However, the trend is only vaguely visible.

Finally, let us look at the distribution of random effect predictions over domains. They are visualized in Fig.7. The bars of the histogram correspond to the probability density of the mode predictors in the respective interval of the support. The red line is

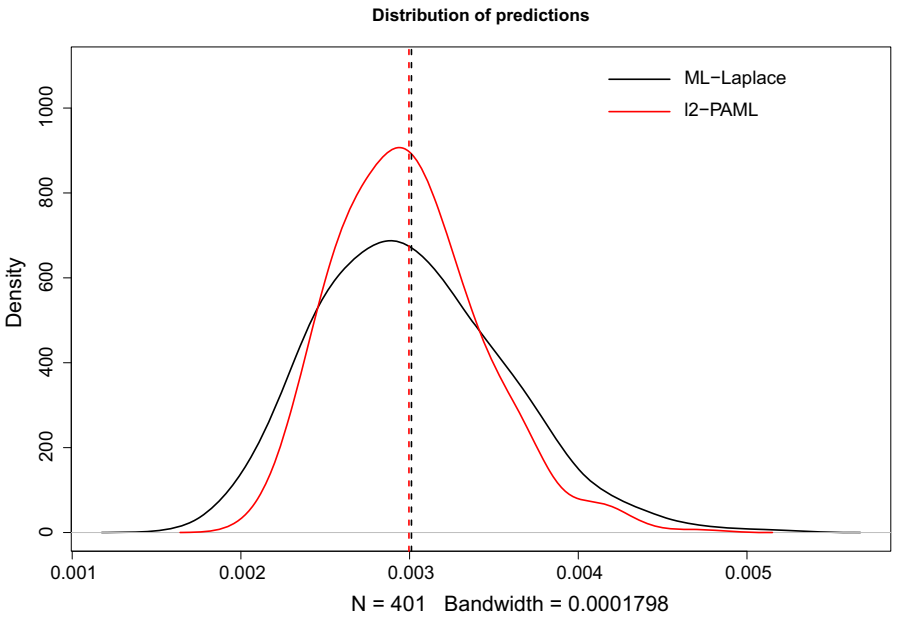

Fig. 4 Comparision of the EBPs

the result of a kernel density estimation over their realized values. We observe that the distribution is very close to normal. This is in line with the theoretical developments from Sect. 2.1. Overall, it can be concluded that the ℓ_2-PAML approach in the area-level logit mixed model was a sensible choice for the considered application.

6 Conclusion

Regional prevalence estimation is an important issue to monitor the health of the population and for planning capacities of a health care system. A good covariate on the prevalence of a disease can be typically obtained from medical treatment records such as the DRG-Statistics in Germany. We proposed a new small area estimator for regional prevalence that copes with two major issues in this context. First, typically health surveys do not have a large sample and the sample size is mainly dedicated to allow for the estimation of national figures. Within regional entities, therefore, the sample size is very small. Applying classical design based or model assisted estimators on these small sample sizes leads to very high standard errors for many regions. Our small area estimator is capable of overcoming this issue by using a model based approach. The second problem we tackle is, that the best covariates at hand, typically have high correlations between each other. This leads to numerical problems inhibiting the exploitation of these covariates. To overcome this problem we propose to use a ℓ_2-penalization approach. This leads to the need for revising the parameter estimation procedure and to adapt it to the new requirements. We provide therefore

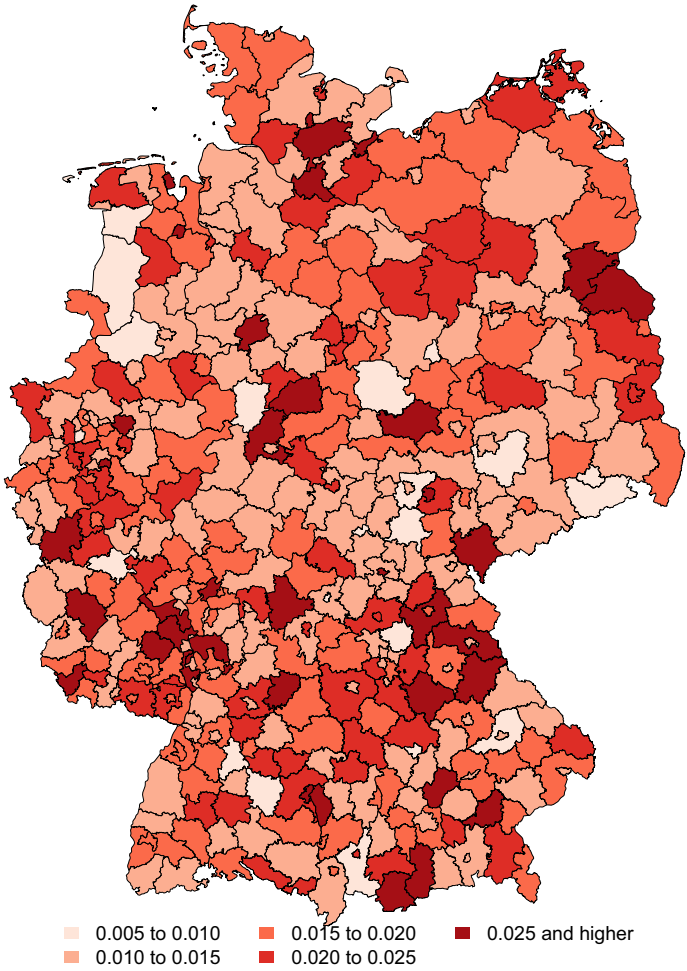

Fig. 5 Results of RRMSE estimation

a novel Laplace approximation to a logit mixed model with ℓ_2 regularization. This estimation procedure is applicable for other purposes such as classical logit mixed model estimation with ℓ_2-penalization.

The prevalence estimation maps of Sect. 5 show some clusters of small areas with high or low prevalence. This fact indicates that modeling spatial correlation by introducing, for example, simultaneous autoregressive random effects, might benefit the final predictions. Combining this additional generalization with the robust penalized approach is thus desirable. However, it is not an easy theoretical task and deserves future independent research. In a Monte Carlo simulation study we show that the proposed estimation approach ℓ_2-PAML yields stable parameter estimates even under strong correlations of the covariates. This simulation results underpin the theoretical arguments. Finally, we applied this newly derived estimator to the prediction of

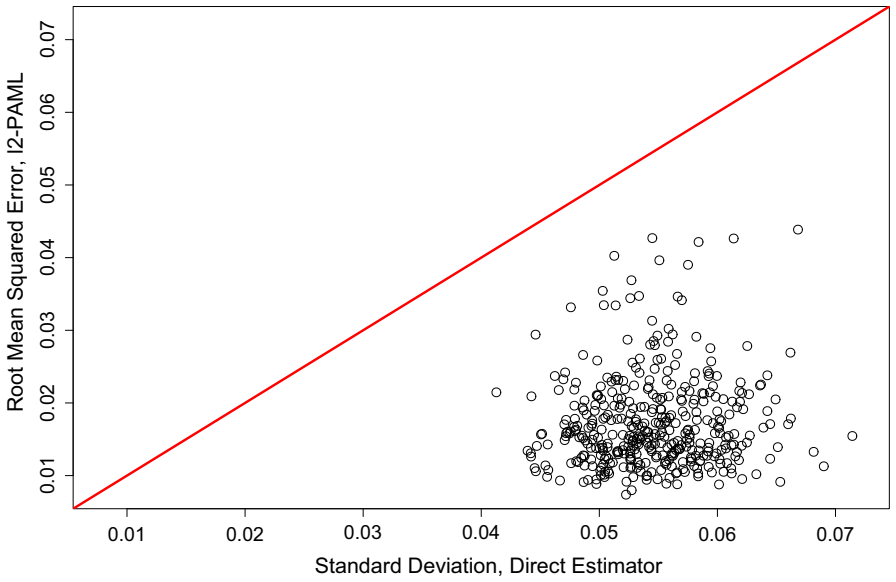

Fig. 6 Comparision of estimation uncertainty

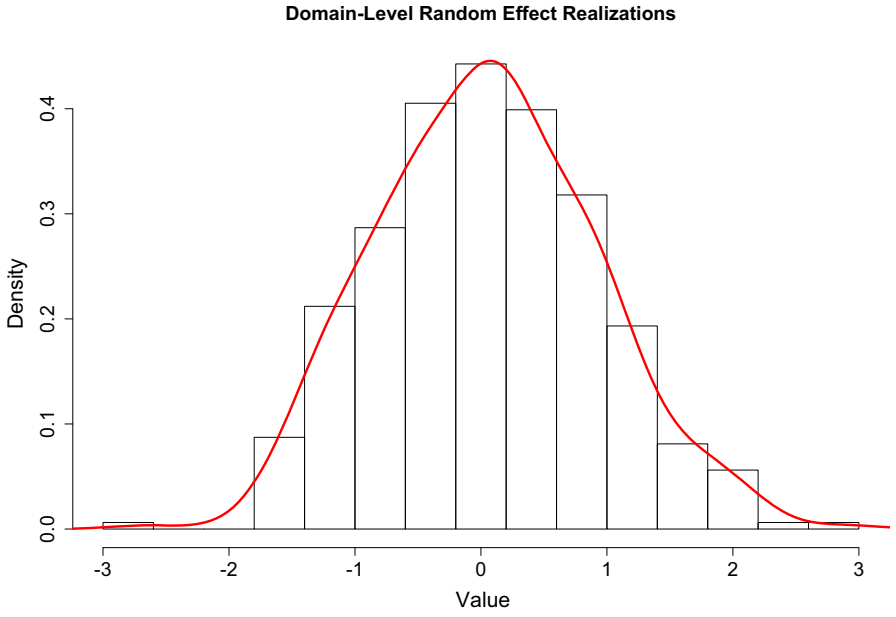

Fig. 7 Distribution of random effect mode predictions

district-level multiple sclerosis prevalence and obtained estimates with a considerably low root mean squared error. Hence, we recommend using our new approach for the regional prevalence estimation.

Funding Open Access funding enabled and organized by Projekt DEAL. This research is supported by the Spanish grant PGC2018-096840-B-I00, by the grant "Algorithmic Optimization (ALOP) - Graduate School 2126" funded by the German Research Foundation, as well as the research project "Gesundheitsatlas" funded by the Scientific Institute of the German Public Health Insurance Company.

Data availability The demographic data as well as the DRG-Statistic data used in this study are available on request from the German Federal Statistical Office. The health insurance records are property of the German Public Health Insurance Company and subject to special privacy restrictions under the national law. They are not available for data sharing.

Declarations

Conflict of interests The authors declare they have no competing interests.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions Automatic Control 19(6):716–723

AOK Bundesverband (2018) Zahlen und Fakten 2018 mit zusätzlichen Grafiken zur Pflegeversicherung. https://aok-bv.de/imperia/md/aokbv/aok/zahlen/zuf_2018_ppt_final.pdf

Berg, EJ (2010) *A small area procedure for estimating population counts* doctoralthesis, Iowa State University

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn 13:281–305

Boubeta M, Lombardía MJ, Morales D (2016) Empirical best prediction under area-level poisson mixed models. TEST 25:548–569

Boubeta M, Lombardía MJ, Morales D (2017) Poisson mixed models for studying the poverty in small areas. Comput Stat Data Anal 107:32–47

Branscum AJ, Hanson TE, Gardner IA (2008) Bayesian non-parametric models for regional prevalence estimation. J Appl Stat 35(5):567–582

Breitkreutz J, Bröckner G, Burgard JP, Krause J, Mönnich R, Schröder H, Schössel K (2019) Estimation of regional diabetes type 2 prevalence in the german population using routine data. AStA Wirtschafts- und Sozialstatistisches Archiv 13(1):35–72

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88(421):9–25

Burgard, JP (2015) *Evaluation of small area techniques for applications in official statistics* doctoralthesis, Universität Trier

Burgard JP, Krause J, Münnich R (2019) Adjusting selection bias in german health insurance records for regional prevalence estimation. Popul Health Metrics 17(10):1–13

Cessie SL, Houwelingen JCV (1992) Ridge estimators in logistic regression. J R Stat Soc Series C (Appl Stat) 41(1):191–201

Chambers R, Dreassi E, Salvati N (2014) Disease mapping via negative binomial regression m-quantiles. Stat Med 33:4805–4824
Chambers R, Salvati N, Tzavidis N (2016) Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the uk. J R Stat Assoc Series A (Stat Soc) 179(2):453–479
Chen S, Lahiri P (2012) Inferences on small area proportions. J Indian Soc Agric Stat 66:121–124
Chicco D (2017) Ten quick tips for machine learning in computational biology. BioData Min 10(35):1–17
Craven P, Wahba G (1979) Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik 31:377–403
Demidenko E (2013) Mixed models: theory and applications with R. Wiley, Hoboken
Dreassi E, Ranalli MG, Salvati N (2014) Semiparametric m-quantile regression for count data. Stat Methods Med Res 23:591–610
Erciulescu, AL and W A Fuller (2013) Small area prediction of the mean of a binomial random variable *JSM Proceedings - Survey Research Methods Section*, 855–863
Ghosh M, Kim D, Sinha K, Maiti T, Katzoff M, Parsons VL (2009) Hierarchical and empirical bayes small domain estimation and proportion of persons without health insurance for minority subpopulations. Surv Methodol 35:53–66
González-Manteiga W, Lombardía MJ, Molina I, Morales D, Santamaría L (2007) Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. Comput Stat Data Anal 51(5):2720–2733
Hobza T, Morales D (2016) Empirical best prediction under unit-level logit mixed models. J Off Stat 32(3):661–692
Hobza T, Morales D, Santamaría L (2018) Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. TEST 27:270–294
Hoerl A, Kennard R (1970) Ridge regression: biased estimation for nonorthogonal problems. Techometrics 12(1):55–67
Jiang J (2003) Empirical best prediction for small-area inference based on generalized linear mixed models. J Stat Plan Inference 111:117–127
Jiang J (2007) Linear and generalized linear mixed models and their application. Springer, New York
Jiang J, Lahiri P (2001) Empirical best prediction for small area inference with binary data. Annal Inst Stat Math 53:217–243
Liu B, Lahiri P (2017) Adaptive hierarchical bayes estimation of small area proportions. Calcutta Stat Assoc Bulletin 69(2):150–164
Long AN, Dagogo-Jack S (2011) The comorbidities of diabetes and hypertension: Mechanisms and approach to target organ protection. J Clinic Hypertens 13(4):244–251
López-Vizcaíno E, Lombardía MJ, Morales D (2013) Multinomial-based small area estimation of labour fource indicators. Stat Model 13(2):153–178
López-Vizcaíno E, Lombardía MJ, Morales D (2015) Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. J R Stat Assoc Series A (Stat Soc) 178(3):535–565
Marino MF, Ranalli MG, Salvati N, Alfò M (2019) Semiparametric empirical best prediction for small area estimation of unemployment indicators. Annal Appl Stat 13(2):1166–1197
Militino AF, Ugarte MD, Goicoa T (2015) Deriving small area estimates from information technology business surveys. J R Stat Assoc Series A (Stat Soc) 178(4):1051–1067
Molina I, Saei A, Lombardía MJ (2007) Small area estimates of labour force participation under a multinomial logit mixed model. J R Stat Soc Series A (Stat Soc) 170(4):975–1000
Müller T (2018) Multiple Sklerose - Zahl der MS-Krankenhat sich in Deutschland verdoppelt ÄrzteZeitung https://www.aerztezeitung.de/Medizin/Warum-es-heute-so-viele-MS-Kranke-gibt-225639.html
Pereira JM, Basto M, da Silva AF (2016) The logistic lasso and ridge regression in predicting corporate failure. Procedia Econ Finance 39:634–641
Petersen G, Wittmann R, Arndt V, Göpffarth D (2014) Epidemiology of multiple sclerosis in germany: regional differences and drug prescription in the claims data of the statutory health insurance. Der Nervenarzt 85(8):990–998
Ranalli MG, Montanari GE, Vicarelli C (2018) Estimation of small area counts with the benchmarking property. METRON 76(3):349–378
Rao JNK, Molina I (2015) Small area estimation (2nd edn) Wiley series in survey methodology. Wiley, Hoboken, New Jersey

Schaefer RL, Roi LD, Wolfe RA (1984) A ridge logistic estimator. Commun Stat Theory Methods 13(1):99–113

Schwarz GE (1978) Estimating the dimension of a model. Annal Stat 6(2):461–464

Statistisches Bundesamt (2016) Demographische Standards Ausgabe 2016 *Statistik und Wissenschaft Band 17*

Statistisches Bundesamt (2017) Fallpauschalenbezogene Krankenhausstatistik (DRG-Statistik). Diagnosen, Prozeduren und Case Mix der vollstationären Patientinnen und Patienten in Krankenhäusern. *Gesundheit Fachserie 12 Reihe 6.4*

Stern S (2014) Estimating local prevalence of mental health problems. Health Serv Outcomes Res Methodol 14:109–155

Tamayo T, Brinks R, Hoyer A, Kuss O, Rathmann W (2016) The prevalence and incidence of diabetes in germany - an analysis of statutory health insurance data on 65 million individuals from the years 2009 and 2010. Deutsches Ärzteblatt Int 113:177–182

Tzavidis N, Ranalli MG, Salvati N, Dreassi E, Chambers R (2015) Robust small area prediction for counts. Stat Methods Med Res 24:373–395

World Health Organization (2018) International classification of diseases for mortality and morbidity statistics (11th revision)

Yamashita T, Yamashita K, Kamimura R (2007) A stepwise AIC method for variable selection in linear regression. Commun Stat Theory Methods 36(13):2395–2403