Discussion Paper

Small Area Estimation of sickness absence based on the Netherlands Working Conditions Survey

Harm Jan Boonstra, Sumonkanti Das and Jan van den Brakel

January 28, 2021

Abstract

The Netherlands Working Conditions Survey is an annual survey that measures, among many other variables, the amount of sickness leave of employees. In the context of the Arbeidsmarkt Zorg en Welzijn (AZW) program we carry out a feasibility study for estimating sickness leave for a detailed breakdown of the AZW subpopulation of employees in the Human Health and Social Work Activities branch. The breakdown is with respect to a subdivision into 28 regions, a subdivision into 14 AZW subbranches (and the non-AZW branch), as well as the combination of both. Especially for the region by subbranch subpopulations the amount of survey data is too small to use direct estimates for this purpose. Therefore a small area estimation approach is developed based on different types of unit-level multilevel models that account for all classifications of interest as well as for selectivity of the survey response with respect to the target population.

1 Introduction

The Netherlands Working Conditions Survey (Nationale Enquête Arbeidsomstandigheden or NEA, in Dutch) is an annual survey that measures (changes in) the working conditions of employees in the Netherlands. The survey is conducted by CBS (Statistics Netherlands) and TNO (Netherlands Organisation for Applied Scientific Research). The surveyed population consists of all employees from age 15 to (and including) 74 years who work in the Netherlands and who are registered as resident of the Netherlands, excluding those living in institutional households. For detailed information about the most recent 2019 NEA survey, in particular about the methodology used, see Hooftman et al. (2020).

One of the topics surveyed by NEA is sickness absence. Within the Arbeidsmarkt Zorg en Welzijn (AZW) program there is a demand for detailed figures about sickness absence regarding the AZW subpopulation of employees in the Human Health and Social Work Activities branch. Desired figures are estimates of several sickness absence measures by region and AZW subbranch. The regional breakdown considered is defined by the so-called RegioPlus regions, a subdivision of the Netherlands in 28 regions. The subbranch classification is a subdivision of branch into 15 subbranches, 14 of which subdivide the AZW branch and the remaining one corresponds to the collection of all non-AZW branches.[1)] The cross-classification of RegioPlus and AZW subbranch therefore has 420 classes.

Regular estimates based on NEA data are computed using the survey weights (Hooftman et al., 2020). This can be done also for subpopulation estimates, where the weighted sums are restricted over the NEA data subset corresponding to each subpopulation. Such direct estimates, however, suffer from high variances in the case that the numbers of observations in (part of) the subpopulations become small. In such a case a

[1)] AZW StatLine divides the Human Health and Social Work activities sector in 10 main branches, two of which have a total of 6 subbranches. For the purpose of this study the lowest level of main and sub-branches is used, which amounts to the 14 classes referred to here as AZW subbranches or simply sub-branches.

model-based estimation methodology, generally known in official statistics as small area estimation (Rao and Molina, 2015), can help to obtain improved estimates. This usually entails using a multilevel model over all domains such that domain estimates also benefit to some extent from similar data in other domains.

As part of the same AZW program a previous feasibility study of small area estimation of position in the job and current education was carried out using data from the Labour Force Survey (de Vries and Michiels, 2019). The same classifiication variables RegioPlus and AZW subbranch have been used in that study. The NEA survey has a much smaller number of annual observations than the Labour Force Survey, so that the need for a small area estimation method is even more urgent for NEA-based estimates at this level of detail.

Statistics Netherlands also conducts a quarterly business survey on job vacancies and sickness absence (Kwartaalstatistiek Vacatures en Ziekteverzuim, KVZ in Dutch). Responding businesses report the percentage absence of their employees. The surveyed population is different from that of NEA not only due to the different reference period but also because the KVZ measures absence for all employees of Dutch businesses, including employees living abroad. A comparison of the published overall figure of the percentage absence based on KVZ nevertheless showed quite good agreement with that based on NEA (Michiels, 2014).

The paper is structured as follows. In Section 2, the survey design of the NEA and the available data sources considered for the small area estimation models are described. The multilevel models considered in this paper to make small area estimates are described in Section 3. Results are presented in Section 4. The paper concludes with a discussion in Section 5.

2 Data sources

2.1 The Netherlands Working Condition Survey

The Netherlands Working Conditions Survey (Nationale Enquête Arbeidsomstandigheden or NEA, in Dutch) is an annual survey that measures (changes in) the working conditions of employees in the Netherlands. The sampling frame for NEA 2019 consists of $7,632,220$ employees. For these persons the sampling frame contains many demographic variables, such as gender, age, ethnicity, region of residence (including province and RegioPlus), degree of urbanisation corresponding to the municipality of residence, etc. Also included are the design variables, i.e. the variables used in the sampling design of NEA, see Hooftman et al. (2020). The sample design of the NEA is based on stratified sampling of employees. The main stratification variable is a subdivision into 42 classes based on industry code. Besides, young employees (at most 24 years old at 1 October 2019) and persons with a non-western migration background are oversampled to compensate for higher non-response rates amongh those groups. The NEA weighting has also been carried out with respect to this sampling frame.

In this project we focus on the NEA 2019 survey data. The NEA 2019 response dataset consists of $n = 58316$ person records. The target variables relating to sickness

absence[2)] are

1 Percentage absence time: the total number of absent days over the last twelve months divided by the total number of workable days.
2 Binary absence: whether an employee has been absent due to sickness in the last twelve months
3 Absence frequency: the number of absent periods in the last twelve months
4 Number of absence days: the number of absent days in the last twelve months
5 Duration of the last absence (not necessarily in the last twelve months): (1-5 days, 5-20 days, 20-210 days, 210 or more days)
6 Work-relatedness of the last absence (not necessarily in the last twelve months): (mainly work-related, partly work-related, not work-related, unknown)

In this report we focus on variables 1-4. The data for the first variable are percentages, for the second variable binary indicators, and for the third and fourth variables counts. Variables 5 and 6 are categorical variables with more than two exclusive classes.

For all 6 absence variables official figures by RegioPlus and by AZW subbranch, but *not* by their full cross-classification, can be found on the Statistics Netherlands outputbase StatLine. However, since annual direct estimates are not sufficiently precise, these figures are averaged over 3 years of NEA data.

Due to item non-response, the number of observations varies per variable. There are 1196 missings for the percentage absence time, 102 missings for binary absence, 1811 missings for the absence frequency and 1108 missings for number of absence days.

2.2 Direct estimates

Direct estimates based on NEA are computed using the NEA weights, which have been derived using a multiplicative weighting method that matches weighted NEA means to unweighted population means for background characterics gender, age, migration background, industry branch, province, degree of urbanisation and educational attainment (Hooftman et al., 2020).

We compute direct estimates and corresponding variance estimates for all target variables and domains (subpopulations) of interest, so that we can compare to the small area estimates discussed later. The direct estimate for the population mean in a certain domain d of interest is computed as

$$\hat{\bar{Y}}_d = \frac{\sum_{i \in s_d} w_i y_i}{\sum_{i \in s_d} w_i}, \tag{1}$$

where y is the absence variable of interest, s_d is the set of employees in domain d for which y is observed, and w_i are NEA weights. Note that the sum in the numerator and denominator runs over the number of observations obtained from the employees in the sample. Due to item-nonresponse the number of observations differs per variable. In this way, the ratio estimator accounts for the item non-response. Corresponding variance estimates are computed as

$$v(\hat{\bar{Y}}_d) = \frac{1}{n_d(n_d - 1)} \sum_{i \in s_d} (y_i - \bar{y}_d)^2, \tag{2}$$

[2)] In the following whenever we speak of absence we mean absence due to sickness

where $\bar{y}_d$ is the mean of y within s_d and n_d is the number of employees in s_d. Note that weights are not used in these variance estimates. A slightly refined variance estimate would include a variance inflation factor due to the variation in weights within each domain, but we have checked that this effect is really small.

AZW subbranch sample sizes range from 29 (Social work) to 1705 (Nursing), and almost 50000 in the non-AZW remainder. The RegioPlus sample sizes are much more balanced: from 787 (Gooi- en Vechtstreek) to 3947 (Haaglanden en Nieuwe Waterweg Noord). For the cross-classification of subbranch and RegioPlus the sample sizes range from 0 to 3947. In particular, 21 out of the 420 domains have zero sample size (and therefore undefined direct estimates and standard errors) and 23 domains have sample size 1 (and therefore undefined standard errors according to (2)). Note that the actual sample sizes can be slightly smaller depending on the number of item-non-responses for each target variable.

2.3 Additional data sources for small area estimation

To enrich the sampling frame with further potentially useful covariates for small area estimation of the absence indicators, data from several registrations have been matched to the sampling frame:

- The Municipal Base Administration (Gemeentelijke Basisregistratie or GBR in Dutch). From this register, household type was added to the sampling frame. We have used a version corresponding to the second quarter of 2019, since this is close to the reference date of the sampling frame (29 March 2019).
- The register of educational attainment over 2018. This register is nearly complete for younger age groups. Most of the data come from educational registrations and another part is derived from Labour Force Survey observations over several years. We have only used the part based on educational registrations. All other units, as well as those that could not be matched to the sampling frame are assigned educational attainment 'unknown'.
- A register of medicine use over 2018. From this registration we have used the number of distinct medicine prescriptions.
- The so-called BIG register. It contains information on medical profession and specialism of registered healthcare workers. The register over 2018 has been used. In the end we have not used the information from this register because the profession information largely overlaps with the AZW subbranch classification and the information about specialism is too sparse when matched to NEA data.
- A register with data on income and social economic class over 2018. From this registration we derive the social economic class (SEC) covariate (see Appendix A)
- The jobs register of the second quarter of 2019. This register contains information on jobs of employees. For employees with multiple jobs we choose the one with most working hours, which is also the job that most NEA questions refer to. From this source several variables are obtained or derived such as income, pay rate, type of contract, working overtime, etc.

All covariates derived from these registrations are categorical. For all registers, and in particular those with reference periods further away from the sampling frame's reference date, it is the case that a certain percentage of sampling frame employees don't match. For these non-matching cases we use the category 'unknown'.

The AZW subbranch classification variable is derived from the industry code and collective labour agreement variables of the jobs register. In this case non-matching

employees are assigned to the remainder non-AZW class, which contains more than 80% of the employees. An exception is the medicine prescription variable; in that case all employees not found in the medicine register are assigned zero medicine prescriptions, because it probably corresponds to the main reason for non-matching.

3 Unit-level models for small area estimation

The best known unit-level model in small area estimation is the Battese-Harter-Fuller model, also known as nested error regression model, or simply as basic unit-level model (Battese et al., 1988; Rao and Molina, 2015). It is a linear multilevel model with a single batch of random intercepts. For the current project we also consider more general unit-level multilevel models that allow for 1) non-continuous unit-level data such as binary or count data and 2) multiple batches of random effects to account for multiple detailed classification variables such as AZW subbranch, RegioPlus and their interaction.

3.1 General unit-level multilevel models

Let y denote one of the target variable vectors, and y_i the observed value for employee i. We denote the length of y by n, which is the number of rows of the NEA dataset minus the small number of missing values due to item-nonresponse, which differs between the target variables. Let X be an $n \times p$ matrix of covariates selected for inclusion in the model. The multilevel models considered take the generalized linear additive form

$$\begin{aligned} y_i &\overset{\text{ind}}{\sim} f(\mu_i, \phi) \\ g(\mu_i) = \eta_i &\equiv X_i\beta + \sum_{\alpha} Z_i^{(\alpha)} v^{(\alpha)}, \end{aligned} \tag{3}$$

where f is a probability distribution depending on a vector of mean parameters μ and an optional scale or dispersion parameter ϕ, and g is a link function that links the mean vector to the linear predictor η. The latter is defined in terms of the covariate matrix X, with X_i denoting its ith row, and associated regression or fixed effects β, as well as a set of random effect design matrices $Z^{(\alpha)}$ of dimension $n \times q^{(\alpha)}$ and corresponding random effect vectors $v^{(\alpha)}$ of size $q^{(\alpha)}$. Here α runs over the different random effect terms used in the model. In most of the models considered we use three random effect terms, one for RegioPlus intercepts, one for AZW subbranch intercepts and one for their interaction.

Several distributions are considered, depending on the target variable. In all cases a linear Gaussian model has been attempted where f denotes a normal distribution with means μ_i and $\phi = \sigma^2$ a variance parameter for the error term in f. In this case the link function used is always the trivial identity function. For the binary absence variable a binomial/Bernoulli model is used with logistic link, i.e. $g(\mu_i) = \log \frac{\mu_i}{1-\mu_i}$. For the count variabes (target variables 3 and 4 as listed in Section 2) we also use a negative binomial distribution, with a logarithmic link function. In that case a dispersion parameter ϕ is allowed to be inferred from the data.

A Bayesian approach of model fitting and prediction is taken. In particular we use Markov Chain Monte Carlo simulation to fit the models, as discussed further in Subsection 3.6. The vector β of fixed effects is assigned a noninformative prior distribution: $p(\beta) \propto 1$. In

the case of a linear Gaussian model the variance parameter is assigned a default noninformative prior: $p(\sigma^2) \propto 1/\sigma^2$. In the case of a negative binomial model the dispersion parameter is assigned a chi-squared distribution with 1 degree of freedom.

The random effect vectors $v^{(\alpha)}$ for different α are assigned independent prior distributions. To describe the general prior for each vector $v^{(\alpha)}$ of random effects, we suppress superscript α from now on. Each random effect vector v is assumed to be distributed as

$$v \sim N(0, A \otimes V)\,, \tag{4}$$

where V and A are $d \times d$ and $l \times l$ covariance matrices, respectively, and $A \otimes V$ denotes the Kronecker product of A with V. The total length of v is $q = dl$, and these coefficients may be thought of as corresponding to d effects allowed to vary over l levels of a factor variable, e.g. intercepts ($d = 1$) varying over subbranch ($l = 15$). The covariance matrix A describes the covariance structure among the levels of the factor variable, and is assumed to be known. Instead of covariance matrices, precision matrices $Q_A = A^{-1}$ are actually used, because of computational efficiency (Rue and Held, 2005). The covariance matrix V for the d varying effects is parameterized in one of three different ways:

- an unstructured, i.e. fully parameterized covariance matrix
- a diagonal matrix with unequal diagonal elements
- a diagonal matrix with equal diagonal elements

The following priors are used for the parameters in the covariance matrix V:

- In the case of an unstructured covariance matrix the scaled-inverse Wishart prior is used as proposed in O'Malley and Zaslavsky (2008) and recommended by Gelman and Hill (2007).
- In the case of a diagonal matrix with equal or unequal diagonal elements, half-Cauchy priors are used for the standard deviations. Gelman (2006) demonstrates that these priors are better default priors than the more common inverse gamma priors for random effects' variance parameters.

3.2 Linear model

None of the absence variables conforms to a normal distribution. All four target variables considered are discrete, even the absence percentage since it is defined as the number of workable days divided by the number of workable days in the last twelve months. Besides, all variables are bounded below by 0 and above by 1 (or 100% on a percentage scale) or by the number of workable days. The absence frequency actually has an apparent cut-off at 40 absence periods.

Nevertheless, a linear Gaussian multilevel model is a convenient base model to fit, and sometimes works surprisingly well for the purpose of predicting population totals or means, such as is the case for small area estimation. For example, Boonstra et al. (2007) conducted a simulation study based on unemployment data from the Dutch Labour Force Survey, from which they conclude that for the task of estimating municipal unemployment fractions a linear multilevel model performs similarly to a logistic multilevel model tailored to the binary unemployment data.

For the linear model, equation (3) becomes, in vector notation,

$$y = X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)} + \epsilon \quad \text{with} \quad \epsilon \sim N(0, \sigma^2 I_n)\,, \tag{5}$$

where I_n denotes the n-dimensional identity matrix.

3.3 Binomial model

The second target variable, binary absence, is a categorical variable with just two classes. For such variables a binomial distribution specialized to binary data, known also as Bernoulli distribution, is the most obvious distribution to use. The link function, linking the distribution's mean, i.e. the probability of a 'success', to the linear predictor, is commonly taken to be the inverse of the logistic function. The resulting model can be viewed as a multilevel generalization of logistic regression.

In this case, equation (3) becomes,

$$\begin{aligned} y_i &\overset{\text{ind}}{\sim} Be(p_i) \\ \log \frac{p_i}{1-p_i} &= X_i\beta + \sum_{\alpha} Z_i^{(\alpha)} v^{(\alpha)}, \end{aligned} \tag{6}$$

where $Be(p_i)$ denotes the Bernoulli distribution with parameter p_i, i.e. $y_i = 1$ with probability p_i and $y_i = 0$ with probability $1 - p_i$.

3.4 Two-part model for zero and non-zero values

The four target variables on absence contain many zeros. This is natural for the binary absence variable modeled using a binomial model, but especially for the percentage absence the large number of zero elements in y means that a linear Gaussian model description becomes less than ideal. The percentage is actually zero in more than half of the cases. For such so-called zero-inflated data, two-part models can be a better choice, see e.g. Pfeffermann et al. (2008); Chandra and Sud (2012); Krieg et al. (2016) for applications of such models to small area estimation. Such models split the description of the variable of interest into two parts, a model for the binary variable of being zero or not, and a conditional model describing the distribution of the nonzero values.

Here such two-part models are considered for the percentage absence variable. For this purpose, write the target variable as $y_i = \delta_i y_i^*$ where δ_i is 0 if $y_i = 0$ and 1 otherwise, and y_i^* denotes the positive value, in case $\delta_i = 1$. Note that the value of y_i^* is irrelevant if $\delta_i = 0$. Both components are then separately modeled using a model of the form (3). In particular, we use a logistic-binomial model for δ_i and a normal linear model for y_i^*. So,

$$\begin{aligned} \delta_i &\overset{\text{ind}}{\sim} Be(p_i) \\ \log \frac{p_i}{1-p_i} &= X_i\beta + \sum_{\alpha} Z_i^{(\alpha)} v^{(\alpha)}, \end{aligned} \tag{7}$$

and

$$y_i^* = X_i\beta_* + \sum_{\alpha} Z_i^{(\alpha)} v_*^{(\alpha)} + \epsilon_{*i} \quad \text{with} \quad \epsilon_* \sim N(0, \sigma_*^2 I_n), \tag{8}$$

where subscript $*$ is used to distinguish the model parameters of the second model from those of the first model. Note that we use the same covariates and random effects for both model parts, which is not necessary in general. It is a practical choice, and it also seems reasonable that covariates predictive of one of the variables are also predictive for the other. Besides, it is important to include variables predictive of δ_i in the model for y^* to avoid selection bias, which could arise because y^* can only be fitted to the subset of data with positive y (Krieg et al., 2016).

Both models are fitted separately to the data. For computing small area estimates both model fits are used for prediction of the respective variables for unobserved population

units, and these predictions are multiplied together to form the prediction for the variable of interest.

3.5 Negative binomial model

For the absence frequency and number of absence days, both linear and negative binomial multilevel models are applied. The latter are generally more suited for count data, especially count data with overdispersion. Another regularly employed distribution family for count data is the Poisson distribution, which has the property that its mean and variance parameters are equal. The negative binomial model is more general in that it allows for a variance that is (much) larger than the mean, i.e. overdispersion. We allow the dispersion parameter r of the negative binomial distribution to be inferred from the data.

In this case, equation (3) becomes,

$$\begin{aligned} y_i &\overset{\text{ind}}{\sim} NBin(r, p_i) \\ \log \frac{p_i}{1-p_i} &= X_i\beta + \sum_{\alpha} Z_i^{(\alpha)} v^{(\alpha)}, \end{aligned} \tag{9}$$

where $NBin(r, p_i)$ denotes the negative binomial distribution with dispersion parameter $r > 0$ and probability parameter p_i, defined by the probability mass function

$$p(y_i|r, p_i) = \binom{y_i + r - 1}{y_i}(1-p_i)^r p_i^{y_i}. \tag{10}$$

For r a positive integer, the negative binomial distribution is the distribution of the number of successes until a certain number r of failures have occurred. Its mean is $\mu_i = \frac{rp_i}{1-p_i}$, and its variance $V(y_i) = \frac{rp_i}{(1-p_i)^2} = \mu_i(1 + \frac{\mu_i}{r})$. Note that smaller r means more overdispersion compared to the Poisson distribution. As r goes to infinity in such a way that the mean approaches a constant value, the distribution approaches the Poisson distribution. For small values of r the negative binomial distribution can fit data with an 'excess' of zeros, so the two-part model is not considered for the count variables.

3.6 Estimation of the multilevel models

The models are fitted using Markov Chain Monte Carlo (MCMC) sampling, in particular the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990). The full conditional posterior distributions used by the Gibbs sampler are all known distributions that are easy to sample from. For the binomial and negative binomial unit-level models we use the data augmentation approach of Polson et al. (2013), in which the binomial likelihood is represented as a scale-mixture of normal distributions. In the negative binomial case the data augmentation approach of Zhou et al. (2012); Zhou and Carin (2015) results in a closed-form full conditional posterior for the dispersion parameter. The MCMC simulations are run in R (R Core Team, 2015) using package `mcmcsae` (Boonstra, 2021).

The Gibbs sampler is run in parallel for three independent chains with randomly generated starting values. In the model building stage 1000 iterations are used, in addition to a 'burn-in' period of 250 iterations (1000 for negative binomial models). This was sufficient for reasonably stable Monte Carlo estimates of the model parameters and trend predictions. For the selected model we use a longer run of 2000 burn-in plus 10000 iterations of which the draws of every fifth iteration are stored. This leaves $3 * 2000 = 6000$ draws to compute estimates and standard errors. The convergence of

the MCMC simulation is assessed using trace and autocorrelation plots as well as the Gelman-Rubin potential scale reduction factor (Gelman and Rubin, 1992), which diagnoses the mixing of the chains. For the longer simulation of the selected model all model parameters and model predictions have potential scale reduction factors below 1.01 and sufficient effective numbers of independent draws.

3.7 Choice of fixed and random effects

To select the most important covariates a quick search has been carried out using fast and simple linear regression fit measures, in particular the (adjusted) R-squared value. Many models of the form (3) have been fitted to the various target variables. For the comparison of models using the same input data and the same distribution we also use the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) and the Widely Applicable Information Criterion or Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2010, 2013).

Model adequacy of these six selected models is evaluated with posterior predictive checks. This implies that replicate data sets, simulated from the posterior predictive distribution are compared with the originally observed data to study systematic discrepancies and to evaluate how well the selected model fits the observed data (Gelman et al., 2004).

It is outside the scope of this project to optimize the set of selected covariates and random effects for each target variable. Instead a practical choice has been made to use the same set of covariates and random effects for each target variable. This choice seems reasonable also because the target variables are quite strongly related.

The model parameters in (3) are separated in fixed and random effects. After extensive examination of different models, the following two fixed effects components are considered in the final models for all target response variables:

$$\begin{aligned} &sex * (ageclass + ethn) + Prov + urban + edu + stratum + \\ &\quad AZW * (sex + ageclass + nonwestern + permanent) \end{aligned} \tag{11}$$

$$\begin{aligned} &sex * (ageclass + ethn) + Prov + urban + edu + stratum + \\ &\quad AZW * (sex + ageclass + nonwestern + permanent) + \\ &\quad hhtype + income + rate + SEC + nmedclass + \\ &\quad contract + overtime + jobtype \end{aligned} \tag{12}$$

See Appendix A for an overview of the covariates used. It is understood that terms like $sex * ageclass$ in (11) and (12) include both main and interaction effects. The variable AZW is a simple indicator variable for whether an employee is working in the AZW branch or not. Interactions of AZW with some important covariates have been included as fixed effects since the AZW sub-population is the population of main interest in this study.

We will refer to models (11) and (12) as the simple and complex covariate models. Both models include, at least approximately, all variables that are used in the NEA sampling design and the NEA weighting scheme. This is important to avoid or limit the overall bias of small area estimates due to different inclusion probabilities or non-response. The complex covariate model contains in addition several predictive variables (such as income, wage rate per hour (rate), number of medications (nmedclass), socioeconomic

status (SEC), types of job and contract) that have been matched from other registrations, as discussed in Section 2.

The overall predictive power of the covariates is, measured in terms of the R-squared or adjusted R-squared linear regression measures, still rather small. For the simple model it ranges from 0.01 to 0.03 depending on the target variable, whereas for the complex covariate model it is approximately double that. From the additional variables in the complex model the medicine use variable contributes most to the increase of R-squared.

For the selection of random effect components the most important considerations are the aggregation levels of interest. For estimation at a particular aggregation level it is desirable to include in the model effects for all underlying classes. In the small area estimation context the number of observations in many of these classes is usually too small to be able to include these effects as fixed effects. Therefore such effects are modelled as random effects. For this application it means that we include random effects for RegioPlus, AZW subbranch and the interaction of RegioPlus and AZW subbranch. A further choice is whether only intercepts or also other covariate effects are allowed to vary over the classes of the aggregation levels. In the case of multiple varying effects there is a choice between scalar, diagonal or full covariance matrix V in (4). In this study we have opted to include only random intercepts at the three levels mentioned. This means that the linear predictor specification in (3) used in all models is

$$\eta_{r[i],b[i]} = X_i\beta + u_{r[i]} + v_{b[i]} + w_{r[i],b[i]}\,, \tag{13}$$

where subscripts r and b are used to denote the region class of RegioPlus and the subbranch, respectively. The notation $r[i]$ can be read as the region in which employee i resides, and analogously for $b[i]$. The random effects are independently normally distributed as $u_r \sim N(0, \sigma_u^2)$, $v_b \sim N(0, \sigma_v^2)$ and $w_{rb} \sim N(0, \sigma_w^2)$. The fixed effects β and design matrix X correspond to either the simple covariate model (11) or the complex one (12).

3.8 Computing small area estimates based on the estimated models

Various models are estimated for each target variable. This results in MCMC simulations for all the model's parameters. Using these simulations, we can subsequently simulate from the posterior predictive distributions for the small area estimands.

Let $\theta_d \equiv \frac{1}{N_d}\sum_{i\in U_d} y_i$ denote a specific domain mean of interest. Here d denotes the domain (e.g. a region, a branch or a combination of region and branch), U_d is the set of all employees in the population of that domain, $N_d = |U_d|$ is its size, and y is one of the target variables. Every MCMC draw s ($s = 1 \ldots S$) from the posterior distribution of the model parameters yields a draw from the posterior predictive distribution for θ_d

$$\theta_d^{(s)} = \frac{1}{N_d}\left(\sum_{i\in s_d} y_i + \sum_{i\in U_d\setminus s_d} y_i^{(s)}\right), \tag{14}$$

where the first term sums the observed values over the set s_d of NEA respondents (excluding item-non-respondents regarding variable y) in domain d, and the second term adds the simulated predictions for all other employees in the population of domain d. The draws $y_i^{(s)}$ are generated according to the distribution f in (3). For example, in

case of the binomial model for binary data,

$$
\begin{aligned}
y_i^{(s)} &\sim Be\left(p_i^{(s)}\right), \\
p_i^{(s)} &= \text{logit}^{-1}\left(X_i\beta^{(s)} + \sum_{\alpha} Z_i^{(\alpha)} v^{(\alpha)(s)}\right),
\end{aligned}
\tag{15}
$$

with $\beta^{(s)}$ and $v^{(\alpha)(s)}$ corresponding to the sth MCMC draw for the model coefficients. Together, the S draws obtained this way for $\theta_d^{(s)}$ form an approximation of its posterior distribution. We use the means of this approximated distribution as point estimates. Standard errors and credible intervals can be computed from this distribution as well. For a two-part model, both variables δ_i and y_i^* are generated in a similar way according to the data distributions used, and their values multipllied so that (16) becomes

$$
\theta_d^{(s)} = \frac{1}{N_d}\left(\sum_{i \in s_d} y_i + \sum_{i \in U_d \setminus s_d} \delta_i^{(s)} y_i^{*(s)}\right), \tag{16}
$$

4 Results

4.1 Results for percentage absence time

The first target variable studied is percentage absence time, i.e. the total number of absent days over the last twelve months divided by the total number of workable days. This is percentage data. We divide by 100 so that the range is between 0 and 1. More than 50% of the values are 0 and there are a few hundred values equal to 1.

Several models of the form (3) have been fitted to the data with fixed and random effects as described in Subsection 3.7. For this variable we also tried the simpler basic unit-level model, which only uses a single random effects term for the cross-classification of RegioPlus and AZW subbranch. So this model is a linear multilevel model with normally distributed errors, identity link, and with linear predictor (13) without the RegioPlus and subbranch random effect terms u and v. This also corresponds to the model used in de Vries and Michiels (2019) for the labour and education related estimates for AZW based on the Dutch Labour Force Survey.

We first compare estimates based on the basic unit-level model, or BHF model, to those based on a linear multilevel model containing all three random effect terms. The covariates used in both models are those of the simple covariate model (11). The differences between the estimates from these models can most clearly be seen at the subbranch level. Figure 4.1 compares the model-based estimates from these two models with the direct estimates at this level. The Figure shows that the lack of (random) subbranch effects in the BHF model causes the small area estimates at this level to be drawn strongly to a common mean value, except for the very large non-AZW domain. Also, the standard errors of BHF model's estimates are most likely underestimated. The estimates based on the model with all three random effects components is much more conservative in the sense that the estimates are still drawn toward a central value, but much less so, and the standard errors are larger. For childcare this standard error is even larger than the direct estimate's standard error.

It can be expected that small area estimates corresponding to a level that is not accounted for in the model become somewhat synthetic. In such a case differences between the classes at that level are only explained by differences in covariates and

differences regarding the other random effect terms. To some extent the random effect term for RegioPlus × subbranch also allows for differences between subbranches but as all these effects are assumed to be independently normally distributed with a common variance parameter they are more limited in explaining marginal differences between subbranches. We conclude that it is better to use a model that includes random effect terms corresponding to all levels of interest, even if the model criteria do not indicate a much better fit of such a model (as is the case here, see Table 4.1 below). In the following all models are assumed to include all three random intercept terms, as in (13).

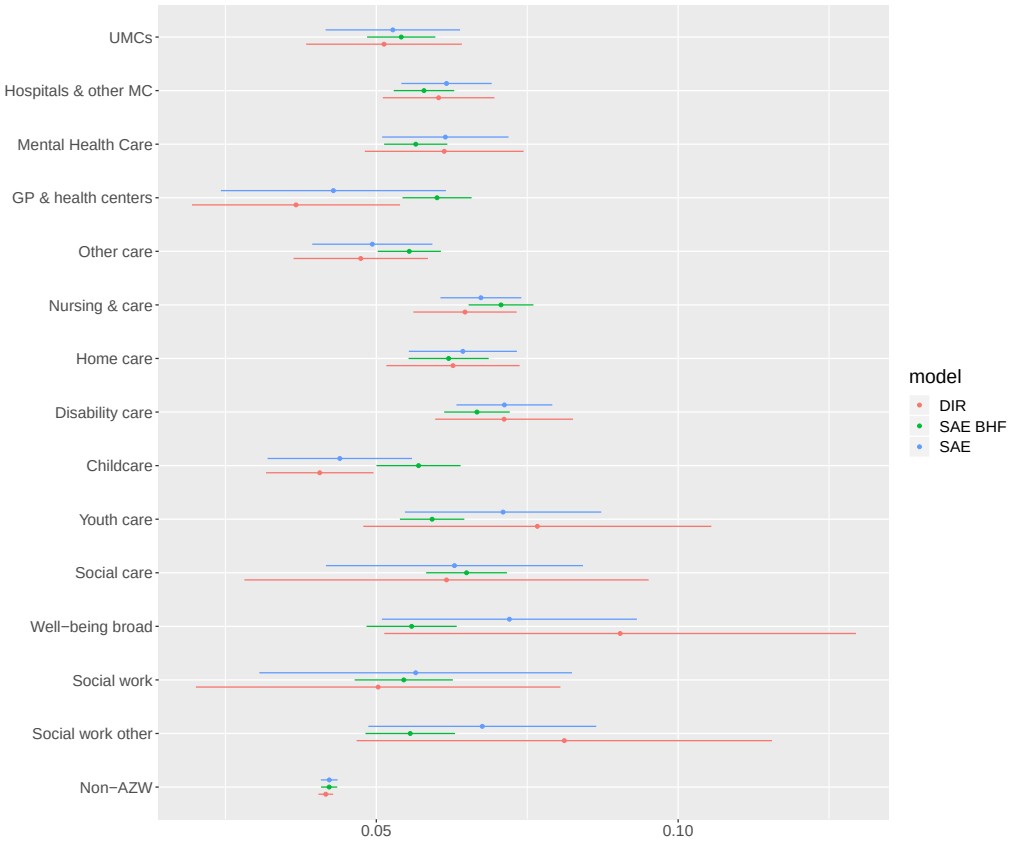

Figure 4.1 Direct (DIR) and model-based estimates with approximate 95% intervals at the subbranch level. The BHF estimates are based on the basic unit-level model with a single random effects component, and the other SAE estimates are based on the same linear multilevel model with additional random intercept components for subbranch and RegioPlus levels.

Table 4.1 lists the model information criteria DIC, WAIC and posterior means of standard deviation parameters for basic unit-level (BHF) model and general linear multilevel models using simple or complex covariate models. Lower values of DIC and WAIC indicate a better overall model fit. These measures also account for the complexity of the model to avoid favouring large models that overfit the data. We note that the DIC and WAIC are estimates and therefore are subject to uncertainty themselves. From this perspective, differences of about 15 units between the information criteria for the BHF and the general linear multilevel models are not very large. The additional random effects' contribution to overall model performance seems modest. Nevertheless, the values for the general linear multilevel models are lower, suggesting that there is a small advantage of including the additional random effects. The more synthetic character of

	BHF_S	BHF_C	LIN_S	LIN_C
DIC	-60832	-63120	-60848	-63134
WAIC	-60819	-63107	-60830	-63120
σ	0.142	0.139	0.142	0.139
σ_u			0.004	0.001
σ_v			0.013	0.014
σ_w	0.003	0.003	0.003	0.002

Table 4.1 Model information criteria DIC, WAIC and posterior means of standard deviation parameters for BHF and general linear multilevel models (LIN) using simple (S) or complex (C) covariate models.

the BHF estimates, however, is a more convincing argument to favour the more general multilevel models.

The differences of DIC/WAIC between the simple and complex covariate models are much larger, as can be seen from Table 4.1, the complex models' measures being more than 2000 units lower. This indicates that the additional covariates used in the complex covariate models substantially improve the overall model fit. The small area estimates based on both models are not very different though. Figure 4.2 shows the estimates by branch. Similar small differences exist for the regional estimates.

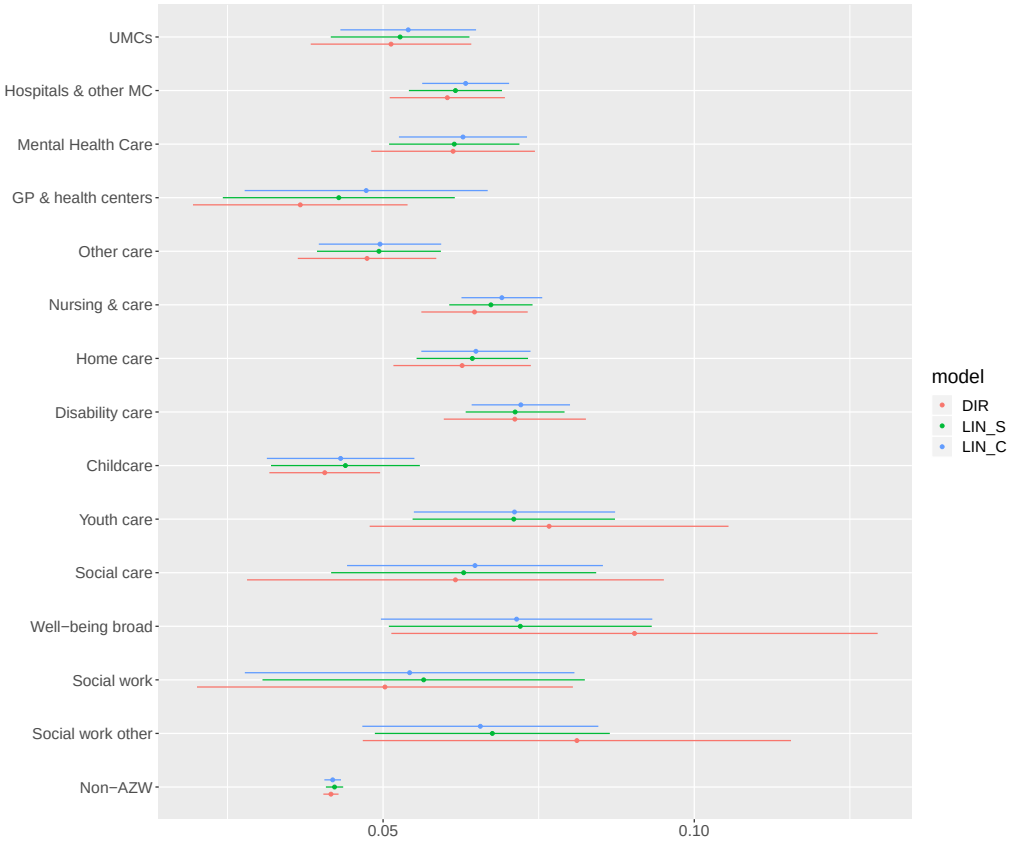

Figure 4.2 Direct (DIR) estimates and estimates based on the linear multilevel models with simple and complex covariate models, for the subbranch level, with approximate 95% intervals.

Table 4.1 also contains point estimates (posterior means) of the models' standard deviation parameters. Here σ denotes the residual standard deviation, the other ones

corresponding to the random intercept components, as in (13). The standard deviation σ_v for subbranch is larger than that of the other random effects, suggesting that differences between subbranches are more pronounced than differences between regions.

Besides the linear models, two-part models have been applied to the percentage variable as well. More than half of the observed absence percentages are zero, and the two-part models take such zero-inflation into account by separately modeling the binary indicator whether the absence percentage is zero or not, see Subsection 3.4. The two-part models considered use a binomial multilevel model for the binary indicator and a linear multilevel model for the positive values. We also tried a log-normal model for the positive values, but that resulted in downward biased small area estimates. The same fixed and random effect components are used in the binomial and linear multilevel models: the three random intercept components as in (13) and either the simple or complex covariate model for both. Figure 4.3 compares the subbranch estimates based on the linear and two-part models, both using the complex covariate model. The differences are not very large, except perhaps for a somewhat higher estimate for GPs and health centers. The standard errors of the two-part model, are on average slightly smaller.

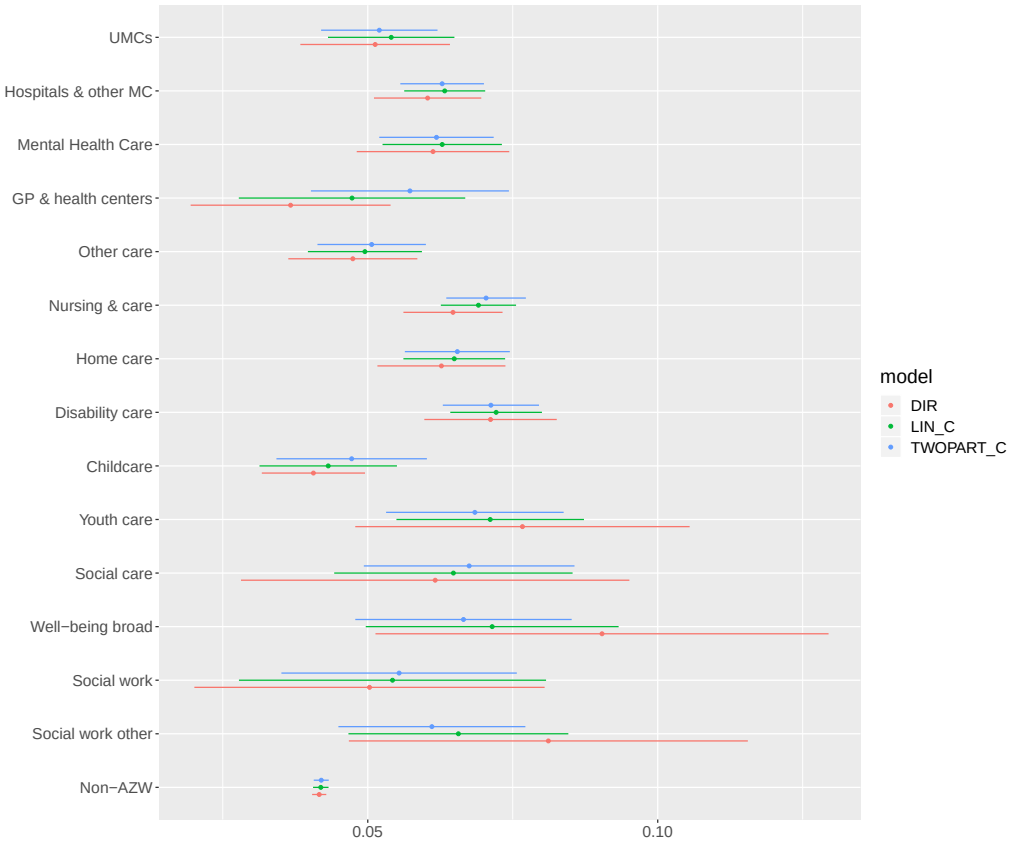

Figure 4.3 Direct (DIR) estimates and estimates based on the linear and two-part multilevel models with complex covariate models, for the subbranch level, with approximate 95% intervals.

It is also possible to extend the two-part model with correlations between the random effects of both model parts (7) and (8), as is done in Pfeffermann et al. (2008). We have tried this for a combination of two linear models for the two components, with a

correlation parameter introduced for all three random effects terms. It turned out that the correlations are quite small, and the small area estimates are very similar to those based on the uncorrelated two-part model. So not only these correlations seem unimportant in our case, but also the choice of logistic-binomial versus normal linear model for the δ variable.

The following figures compare the direct estimates with the estimates based on the two-part model with complex covariate model, for all aggregation levels of interest. Figure 4.4 shows scatterplots of the direct estimates and model-based (SAE) estimates (left), their standard errors (middle) and the coefficients of variation (CV, right), defined as the standard errors divided by the estimates. These are the estimates at the most detailed (RegioPlus by subbranch) level. Note that these figures do not display estimates for which the direct estimate is undefined (due to zero responses) or standard errors/CVs for which the direct standard error is undefined (0 or 1 responses). Looking at the range of the SAE estimates, we see that it is much smaller than the range of the direct estimates. The range of the SAE estimates is much more reasonable. In particular, there are no zero or very small estimates or very high estimates, which in the case of the direct estimates are due to the small sample sizes. The standard errors and CVs of the SAE estimates are much smaller than those of the direct estimates. Most CVs of the model estimates are below 20%. Note that SAE estimates for large domains are more similar to the corresponding direct estimates. This is in particular the case for the non-AWZ branch domains.

Figures 4.5 and 4.6 show the same series of scatterplots at the region and branch levels. Model estimates at the regional level are more similar to the direct estimates due to reasonably large sample sizes in almost all regions. The subbranch classes are much less balanced, so here the model estimates can be different. In particular, the most extreme direct estimates are drawn toward the centre by the model. Standard errors and CVs are almost always smaller at these levels, although the relative gain is not as large as at the most detailed level.

Appendix C shows plots comparing the model estimates to the direct estimates. The first two show these estimates at the regional and subbranch level, and the following four figures show estimates at the most detailed level for a subset of subbranches.

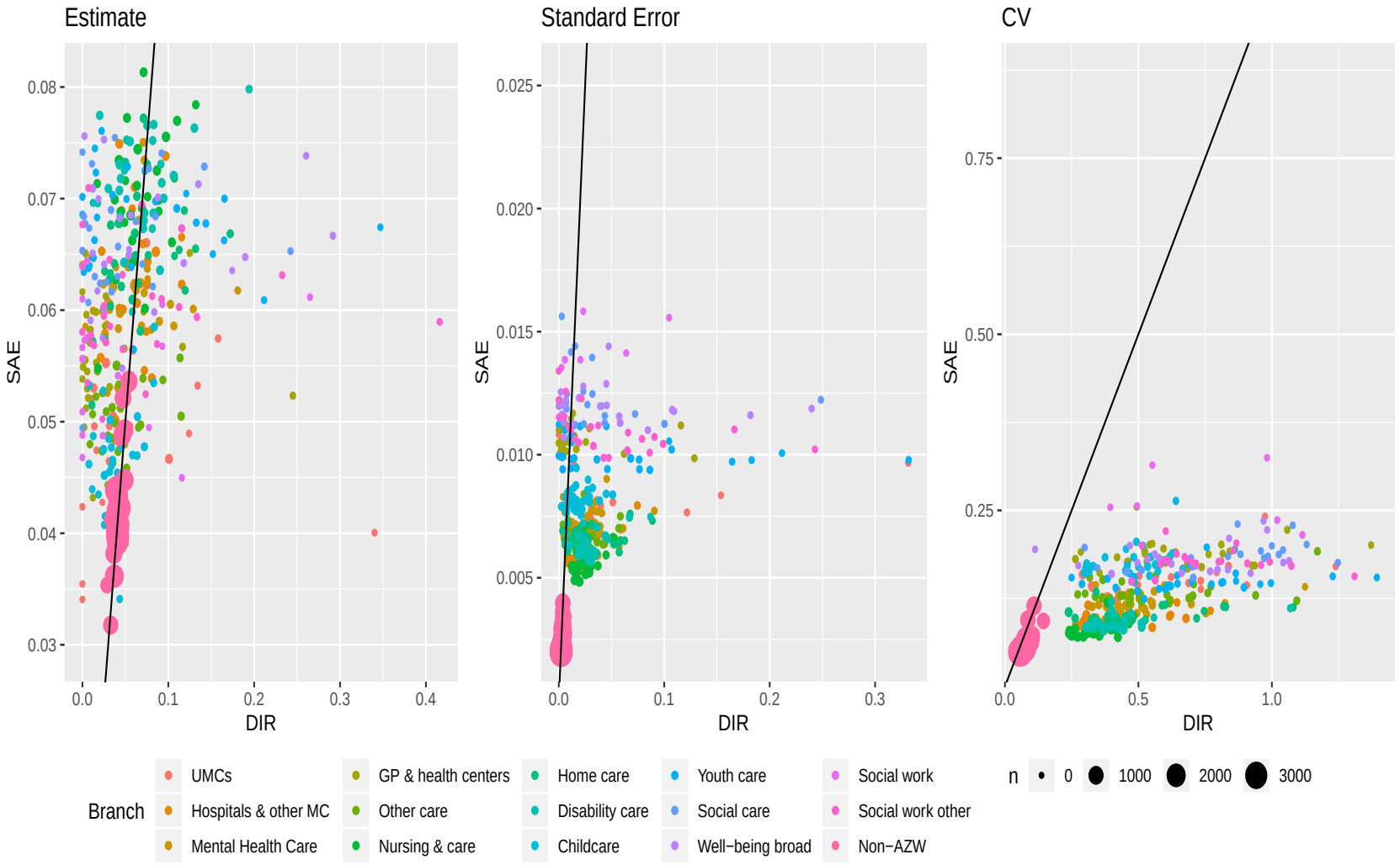

Figure 4.4 Direct (DIR) and model-based (SAE) estimates (two-part multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the percentage absence time due to sickness at the target small domains cross-classified by region and subbranch.

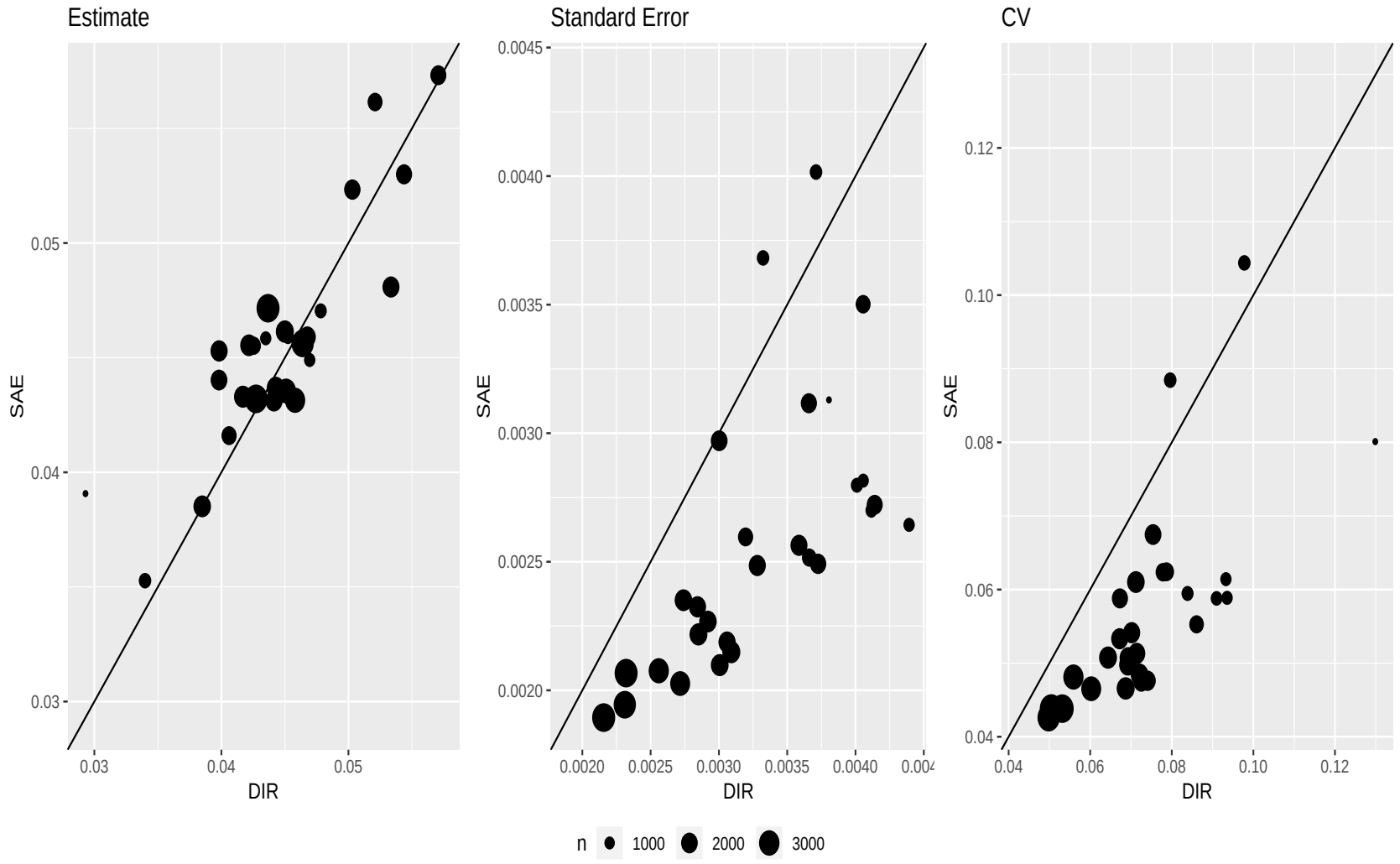

Figure 4.5 Direct (DIR) and model-based (SAE) estimates (two-part multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the percentage absence time due to sickness at the region level.

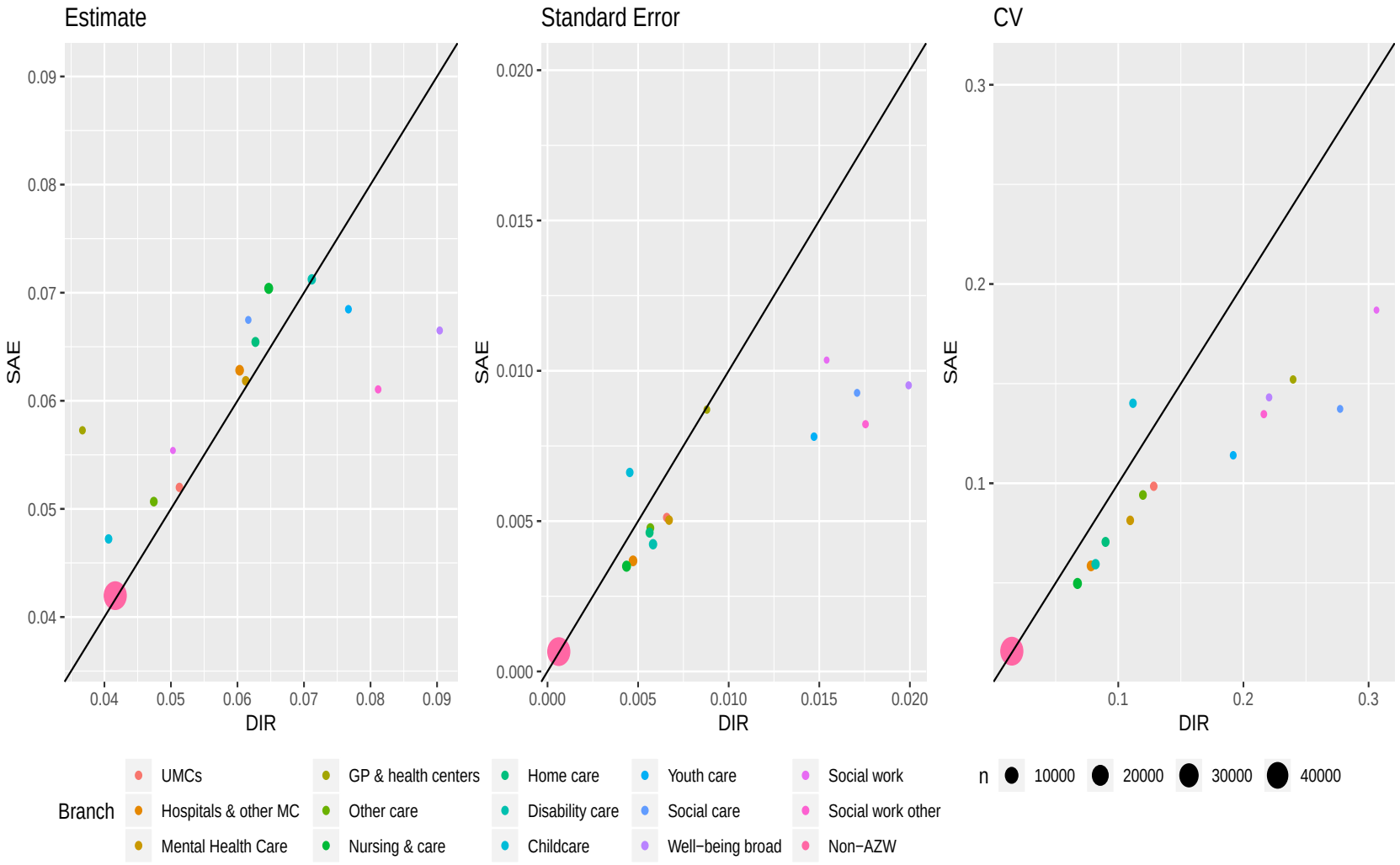

Figure 4.6 Direct (DIR) and model-based (SAE) estimates (two-part multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the percentage absence time due to sickness at the subbranch level.

Finally, we compare the estimates by subbranch with estimates based on the 2019 KVZ as well as the three-year averages over NEA 2017-2019, both as published on StatLine. Note that the reference period for the three-year averages is actually 2018. This illustrates one main disadvantage of taking averages over subsequent years: it delays the figures since the reference period is the middle of the period over which averages are taken. On top of that, evolutions in trends are leveled out. Figure 4.7 shows the comparison of these estimates with direct (NEA 2019) and benchmarked small area estimates based on the two-part model with complex covariate model. KVZ and NEA three-year averages standard errors were not available, so they are displayed only as point estimates. Also, the NEA three-year averages were not available for "Social work" and "Social work other" subbranches (suppressed in the StatLine table). The figure shows that for most subbranches the small area estimates are closer to the KVZ subbranch estimates than are the direct estimates. A notable exception is subbranch "GPs and health centers".

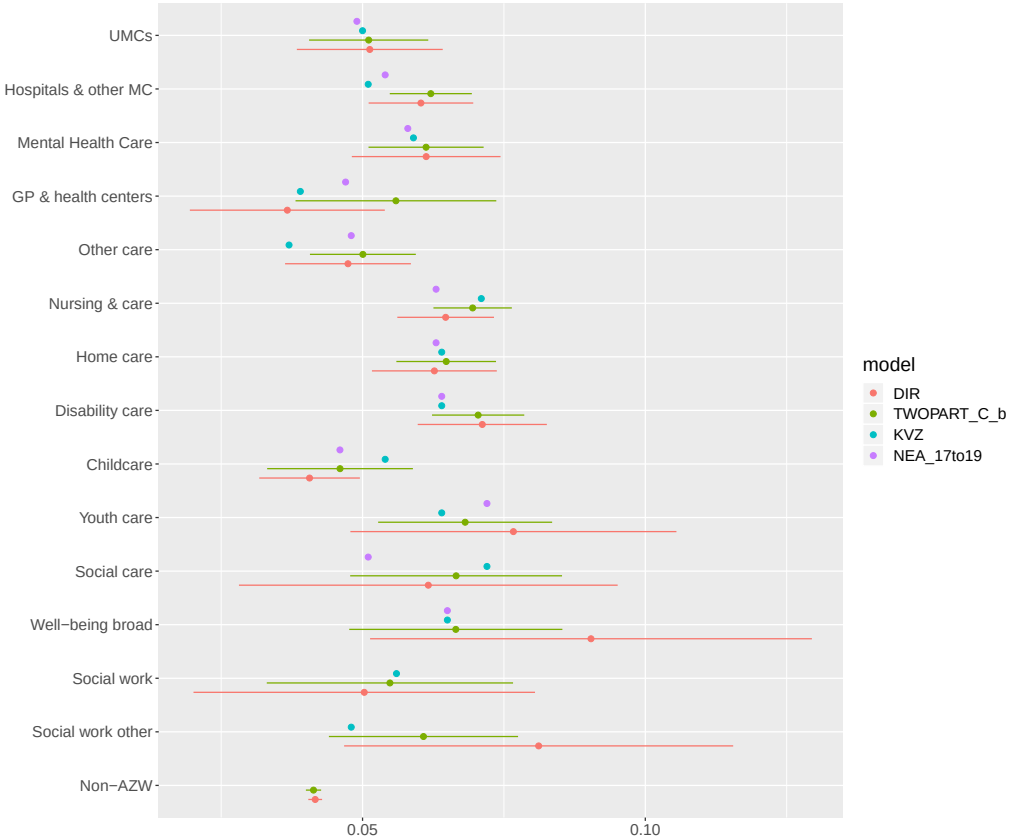

Figure 4.7 Direct NEA 2019 estimates (DIR), benchmarked small area estimates based on the two-part multilevel model with complex covariate model (with approximate 95% intervals), and point estimates based on the KVZ business survey as well as three-year averages based on NEA 2017-2019, all at the subbranch level.

4.2 Results for percentage of employees with absence

The next target variable is the binary indicator of whether an employee has been absent due to sickness in the last twelve months, or not. For this variable we try both linear and binomial multilevel models, both using either the simple or complex covariate model.

Figure 4.8 shows a comparison of the subbranch level estimates between the linear models with simple and complex covariate models. Differences between simple and complex model estimates are mostly small, with some larger differences for GPs and health centers and the non-AZW branch. Standard errors are slightly smaller for the complex covariate model, which is also favoured by the DIC/WAIC criteria (Table 4.2).

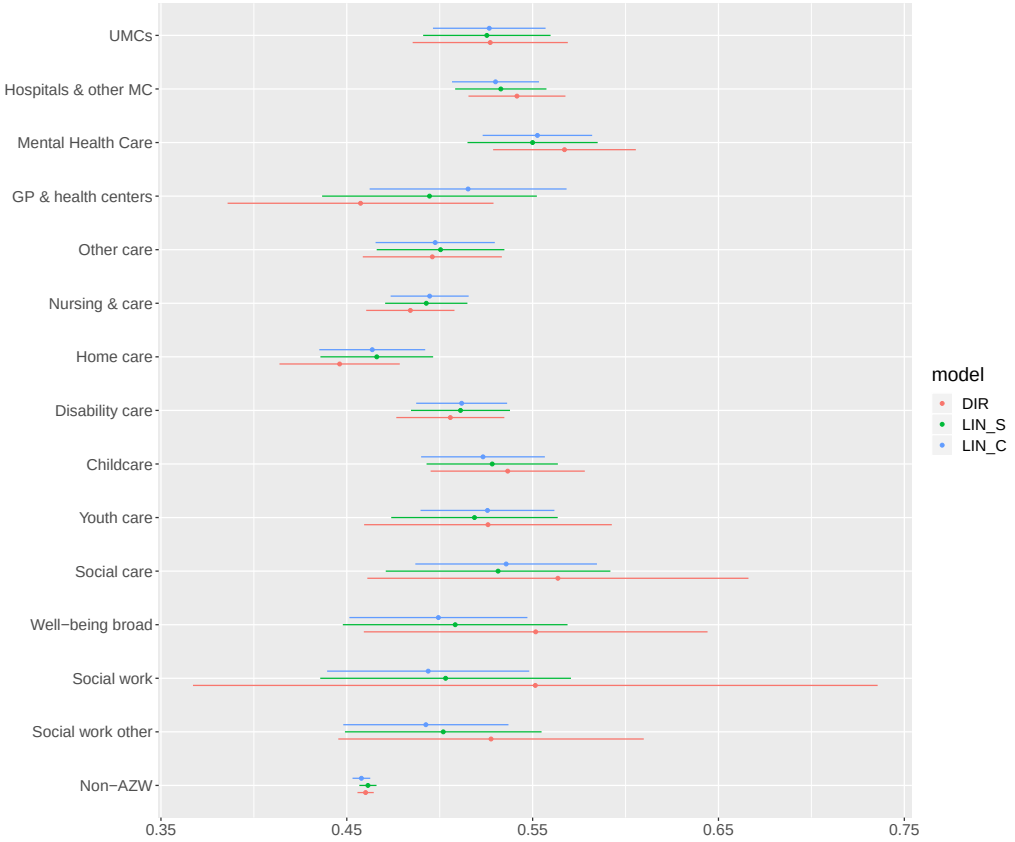

Figure 4.8 Direct (DIR) estimates and estimates based on the linear multilevel models with simple and complex covariate models, for binary absence at the subbranch level, with approximate 95% intervals.

	LIN_S	LIN_C	BIN_S	BIN_C
DIC	82926	80899	79125	77054
WAIC	82926	80898	79125	77055
sigma_u	0.004	0.004	0.018	0.017
sigma_v	0.029	0.020	0.118	0.085
sigma_w	0.005	0.005	0.021	0.019

Table 4.2 Model information criteria DIC, WAIC and posterior means of standard deviation parameters for linear (LIN) and binomial (BIN) multilevel models using simple (S) or complex (C) covariate models. Note that DIC/WAIC are only comparable between models using the same data distribution.

A more natural model for the binary absence variable is a binomial multilevel model with logistic link function. Figure 4.9 shows the estimates based on linear and binomial models at the subbranch level. The differences turn out to be negligible. The same is true for the small area estimates at the other levels of interest. Apparently, for the binary absence variable, with means not far from 0.5 and always far from 0 and 1, it

doesn't matter whether small area estimates are based on the linear or binomial model, even though unit-level predictions based on the binary model are much more appropriate. For variables taking values close to the boundaries of the admissible range (zero or one), larger gains with binomial model might be expected. A practical advantage of using the linear model is that prediction for the population can be done much faster by precomputing domain-level population aggregates of covariate and random effect matrices.

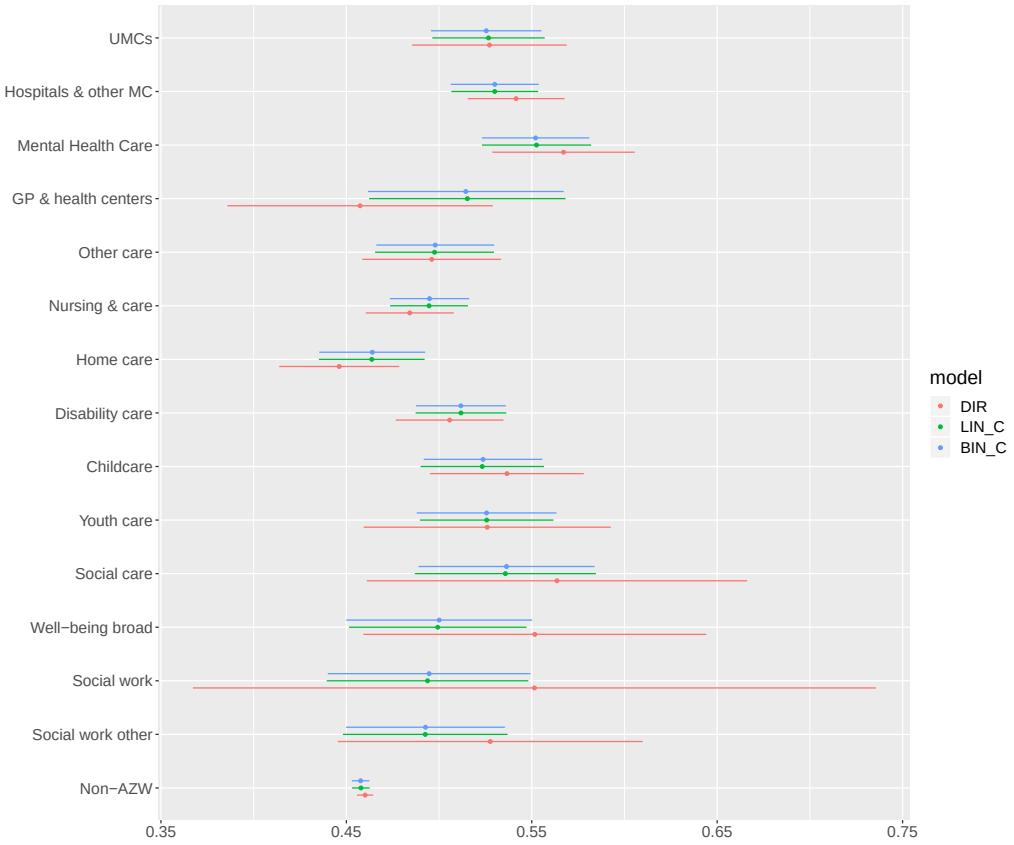

Figure 4.9 Direct (DIR) estimates and estimates based on the linear and binomial multilevel models with complex covariate models, for binary absence at the subbranch level, with approximate 95% intervals.

The next plots compare the direct estimates with the estimates based on the binomial model with complex covariate model, for all aggregation levels of interest. From figure 4.10 it is clear that the range of the SAE estimates is again much smaller and more realistic than the range of the direct estimates. Many direct estimates are even 0 or 1 due to small sample sizes. The standard errors and CVs of the SAE estimates are much smaller than those of the direct estimates. In this case all CVs of the model estimates are below 10%.

Figures 4.11 and 4.12 show the same series of scatterplots at the region and branch levels. Model estimates at the regional level are, again, quite similar to the direct estimates, and at the subbranch level there are a few more pronounced differences.

Appendix C shows additional plots comparing the model estimates to the direct estimates, the first two at the regional and subbranch levels, and the following four at the most detailed level for a subset of subbranches.

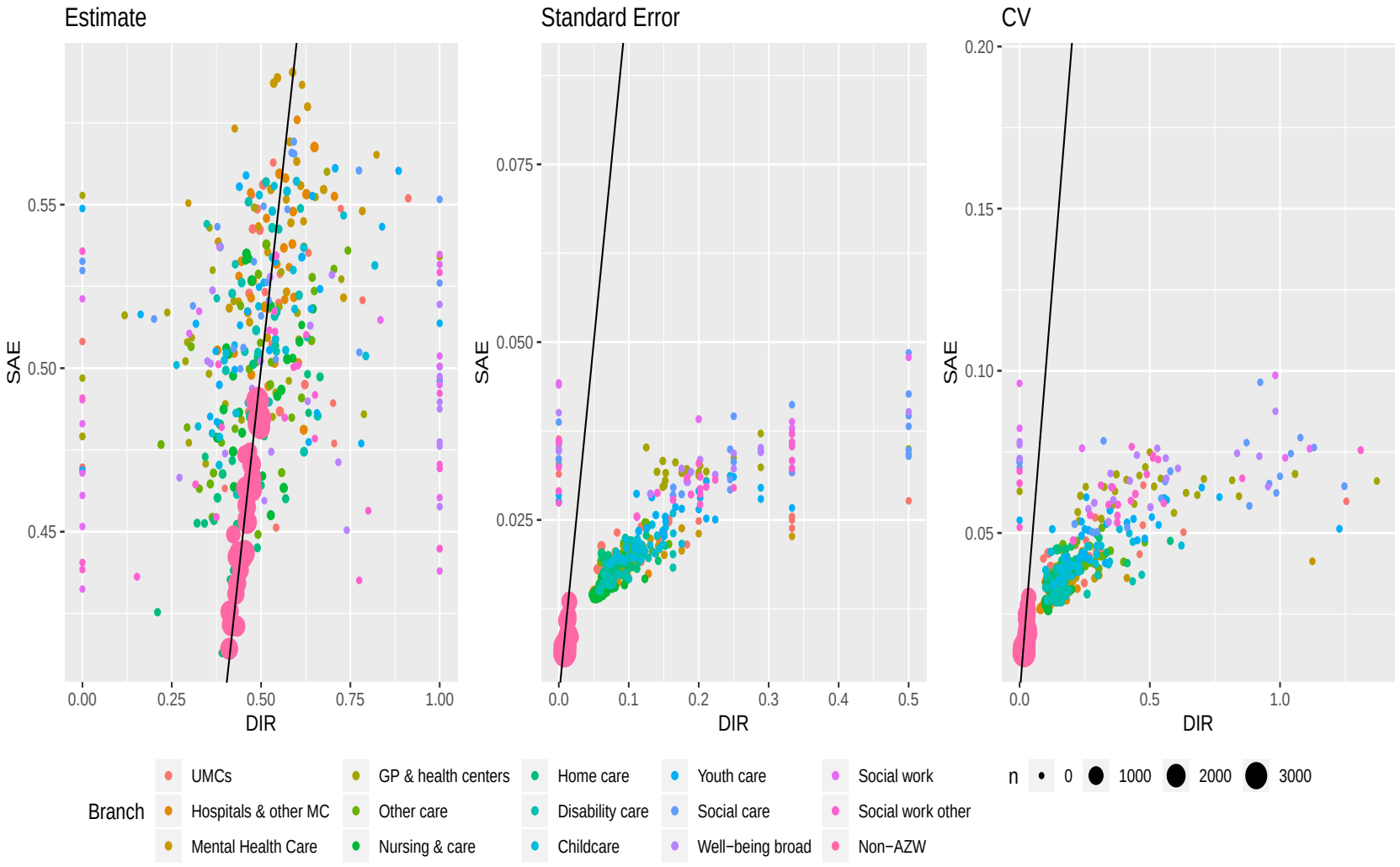

Figure 4.10 Direct (DIR) and model-based (SAE) estimates (binomial multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the percentage of absent employees due to sickness at the target small domains cross-classified by region and subbranch.

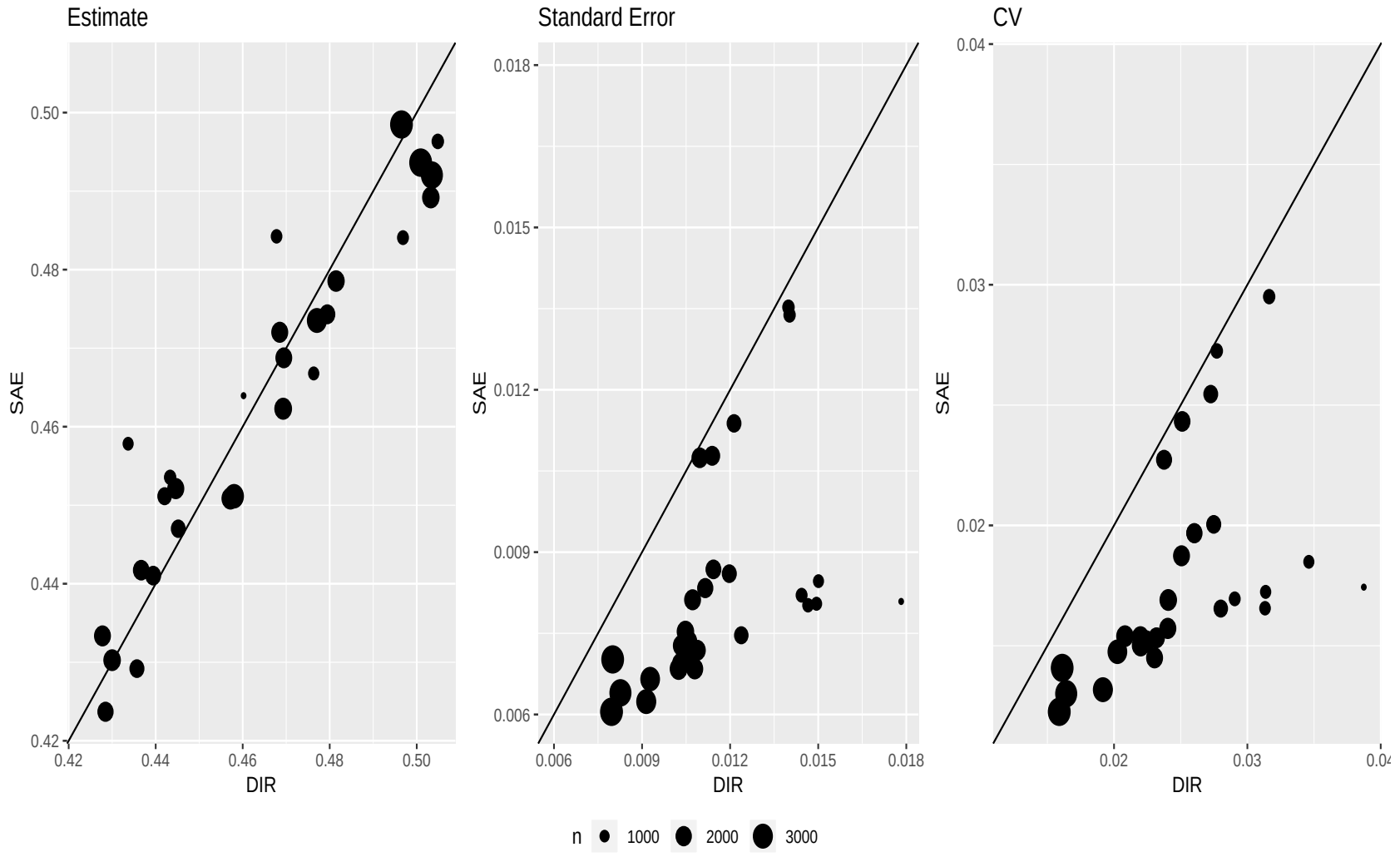

Figure 4.11 Direct (DIR) and model-based (SAE) estimates (binomial multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the percentage of absent employees due to sickness at the region level.

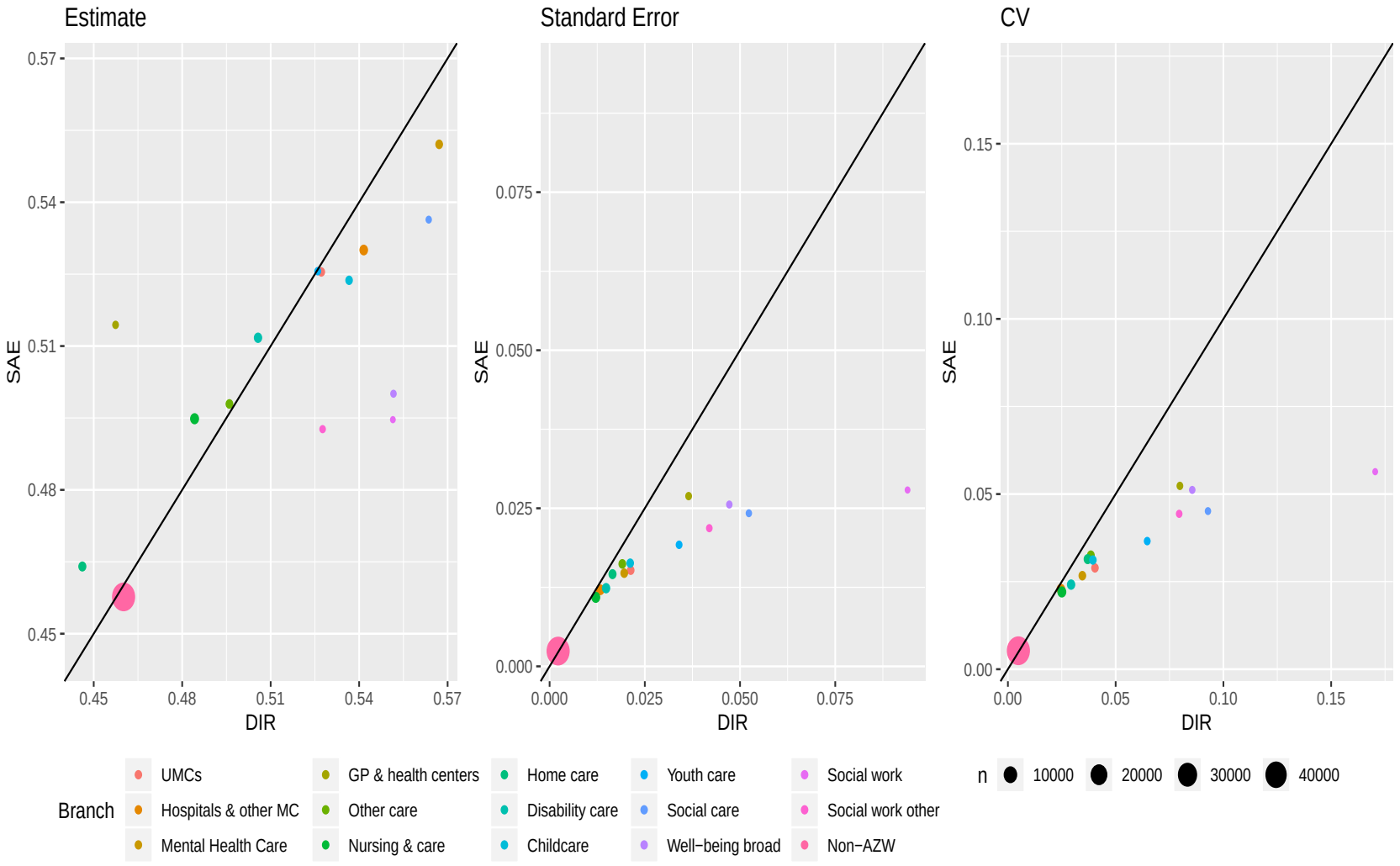

Figure 4.12 Direct (DIR) and model-based (SAE) estimates (binomial multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the percentage of absent employees due to sickness at the subbranch level.

4.3 Results for number of absence periods

The third target variable is the frequency of absent period due to sickness over the last twelve months. This is a count variable with a skewed distribution having more than 50% zero observation. Consequently a Poisson or a negative binomial regression model is expected to portray the actual data structure. Here we develop both linear (5) and negative binomial (9) multilevel models for predicting the mean number of absent periods at the considered domain levels.

The linear multilevel model provides reasonable estimates at the overall, region and branch levels, but the estimates at the detailed domain level tend to overshrink to the average level. Also, the standard errors (SE) based on the linear model appear to be underestimated at the detailed level, see Figures E.2 and E.3 in Appendix E. The posterior predictive checks also confirm how the linear model failed to represent the actual distribution of the count variables, see Figure E.1. As expected from the analysis of the previous two variables, the model with complex fixed effect component provides considerably lower DIC and WAIC values compared to the model with the simple fixed effect component. This is true for both linear and negative binomial multilevel models. See Table 4.3, which also contains (posterior mean) estimates of model parameters (other than coefficients). Note that the DIC and WAIC criteria can only be compared among the two linear or among the two negative binomial models, and not between models using different data distributions.

statistics	LIN_S	LIN_C	NB_S	NB_C
DIC	305089	304192	164696	162505
WAIC	305129	304235	164987	162838
σ	3.60	3.57		
r			0.41	0.45
σ_u	0.03	0.02	0.05	0.05
σ_v	0.10	0.10	0.10	0.12
σ_w	0.04	0.03	0.31	0.32

Table 4.3 Model information criteria DIC, WAIC and posterior means of standard deviation parameters for linear (LIN) and negative binomial (NB) multilevel models using simple (S) or complex (C) covariate models for the number of absent periods.

From here on we use the negative binomial model with complex covariate model to compare to the direct estimates. Figure 4.13 (the left panel) shows the comparison between direct (DIR) and model-based estimates using the negative-binomial model with the complex fixed effect component for the mean number of absent periods for the 420 cross-classified domains of region and subbranches. The middle and right panels show how the SAE estimates improved the accuracy (SE) and the relative accuracy (CV) of the estimates. Note that the range of SAE estimates is much more reasonable than the range of direct estimates At region and subbranch levels, the gain of using the SAE method is also evident from the comparisons of SEs and CVs shown in Figure 4.14 and Figure 4.15. These figures also show that on average the model-based estimates are slightly larger than the DIR estimates. Average discrepancies between model and direct estimates are discussed further in Subsection 4.5.

Point estimates of the mean number of absent periods with approximate 95% interval at region and subbranch levels are shown in Figures 4.16 and 4.17, respectively. At the most detailed level, the gain is evident, especially for domains with zero or a very small

number of observations. See for example Figure E.5 in Appendix E for the “Social work” subbranch where direct estimates are in most cases 0, 1, or missing. For larger subbranches the direct estimates are less extreme, but the gain by small area estimation is still clear, see Figures E.6 and E.7. For subbranch “GP and health centers”, however, the SAE estimates seem systematically higher with sometimes larger uncertainty than the direct estimates, though the sample sizes are small (1-15 in all regions), see Figure E.4.

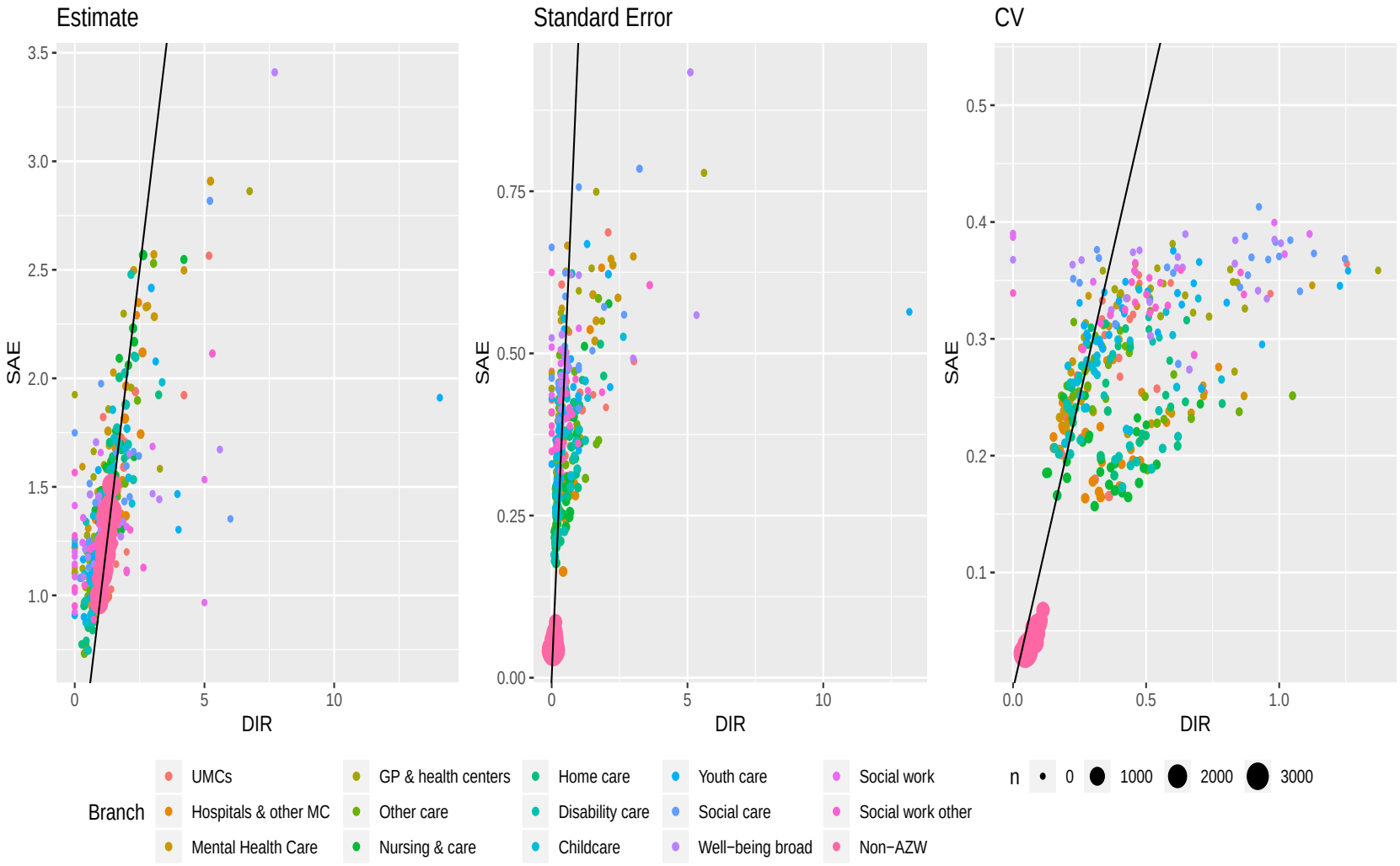

Figure 4.13 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the number of absence periods due to sickness at the target small domains cross-classified by region and subbranch.

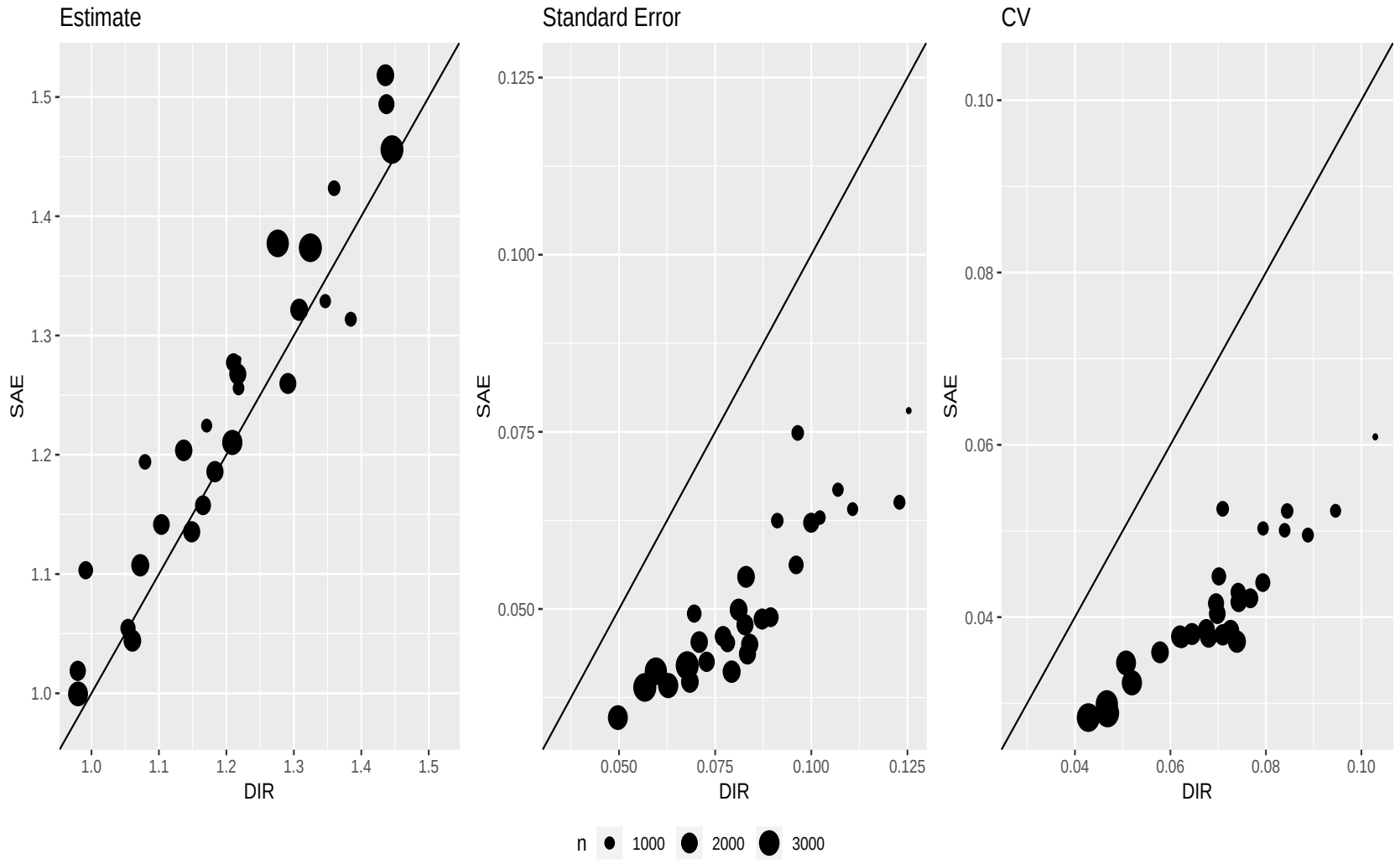

Figure 4.14 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the number of absence periods due to sickness at the region level.

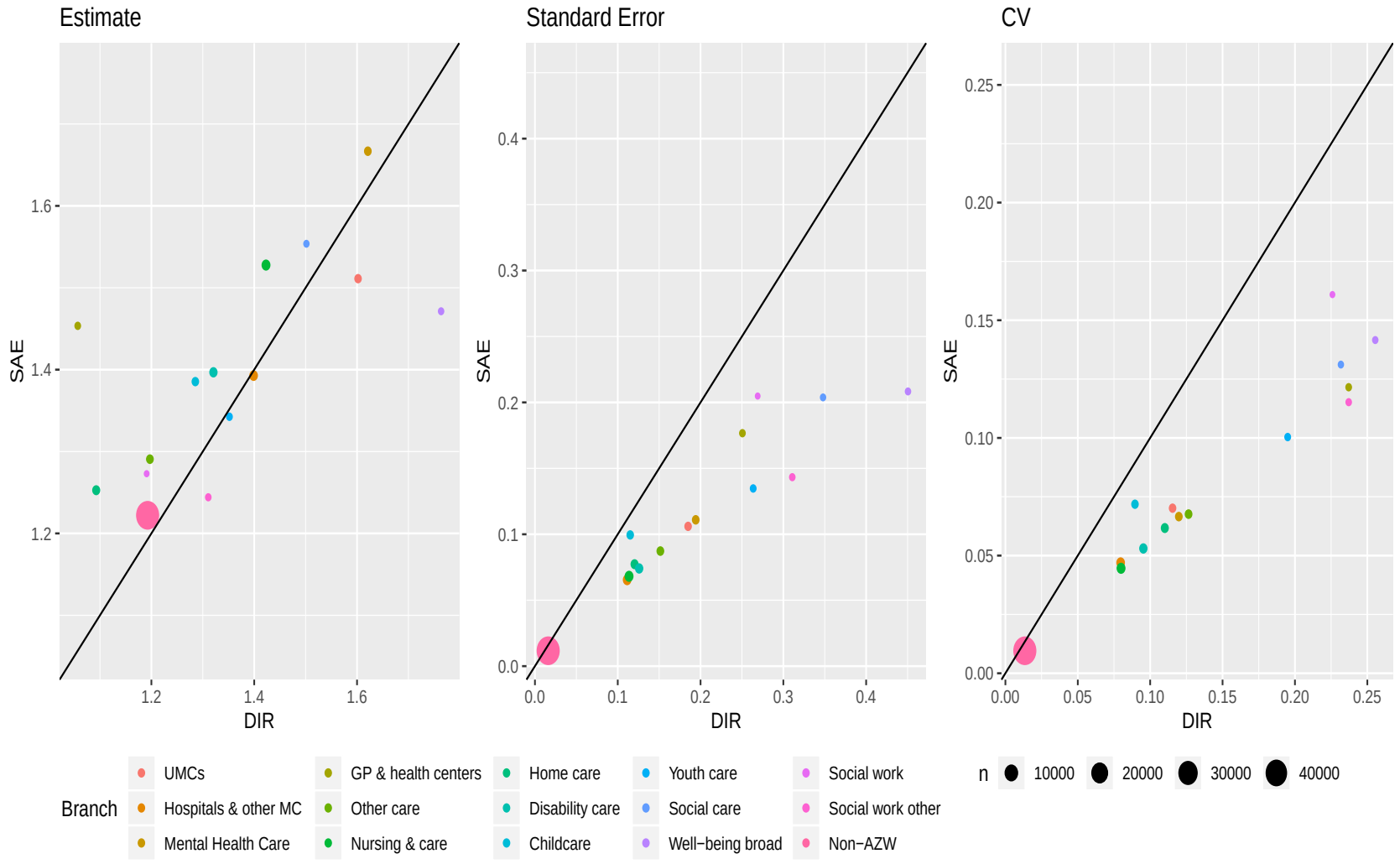

Figure 4.15 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the number of absence periods due to sickness at the subbranch level.

Figure 4.16 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with approximate 95% interval for the number of absence periods due to sickness at the region level.

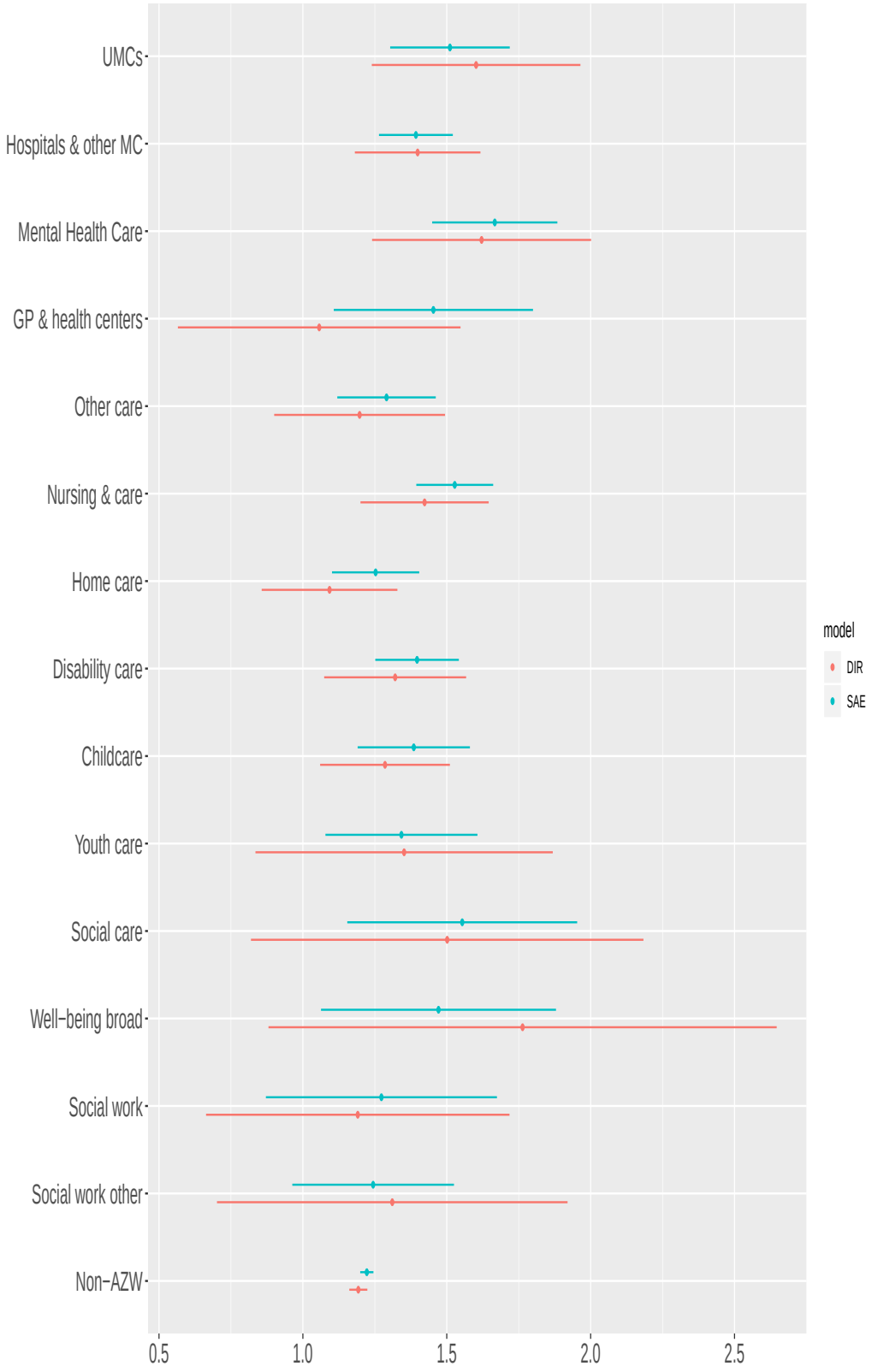

Figure 4.17 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with approximate 95% interval for the number of absence periods due to sickness at the subbranch level.

4.4 Results for the number of absence days

The last target variable is the actual number of absent days during the last 12 months. Like the number of absence periods variable, this variable is a count variable, but with a wider distribution (varying from 0 to 315). See the actual distribution of the number of absence periods and absent days in Figures E.1 and F.1. Both linear and negative binomial multilevel models have been fit to the data and as expected the linear models fail to portray the actual distribution of the number of absent days (Figure F.1). The smaller estimated dispersion parameter of the negative binomial multilevel model (0.16 and 0.17 in the models with simple and complex fixed effect components respectively) indicates that the variable is highly dispersed.

As for the other target variables, the multilevel model with complex fixed effect component has reduced the DIC and WAIC values by more than 2000 units for both linear and negative binomial based models. Also, the variance components are somewhat smaller in the models with complex fixed effects, particularly for the linear multilevel model (see Table F.1). Though the linear multilevel model fails to follow the actual distribution, it provides reasonable estimates at all the target levels except the most detailed level. Figures E.2 and E.3 in Appendix E seem to indicate that the SAE estimates and standard errors based on the negative binomial model are more plausible than those based on the linear model.

A comparison between direct and SAE estimates using the negative-binomial model with the complex fixed effect component of the mean number of absent days for the detailed cross-classified domains, is shown in Figure 4.18. Similar to the number of absent periods, the SAE method provides slightly higher estimates for most of the AZW small domains (the left panel plot) with relatively better accuracy (the right panel CV plot). As before, SAE estimates for the larger domains of the non-AZW subbranch are very similar to those of the direct estimates. Figures 4.19 and 4.20 show that the regional SAE estimates are more precise than those at the subbranch level in terms of SE and CV. The point estimates with 95% CI at region and subbranch levels shown in Figures 4.21 and 4.22 also reveal that the SAE method still brings some modest gain at the regional level. At the detailed level, the gain is evident for most of the domains especially for those with zero or very small sample size. These estimates, however may be somewhat synthetic, since they tend strongly to the average number of absent days, see Figures F.4, F.5 and F.6. A small gain is also observed for the larger detailed non-AZW domains (see Figure F.7). For the domains under the subbranch “GP and health centers”, the performance of the SAE estimates shown in Figure F.4 is very similar as found for the number of absent periods.

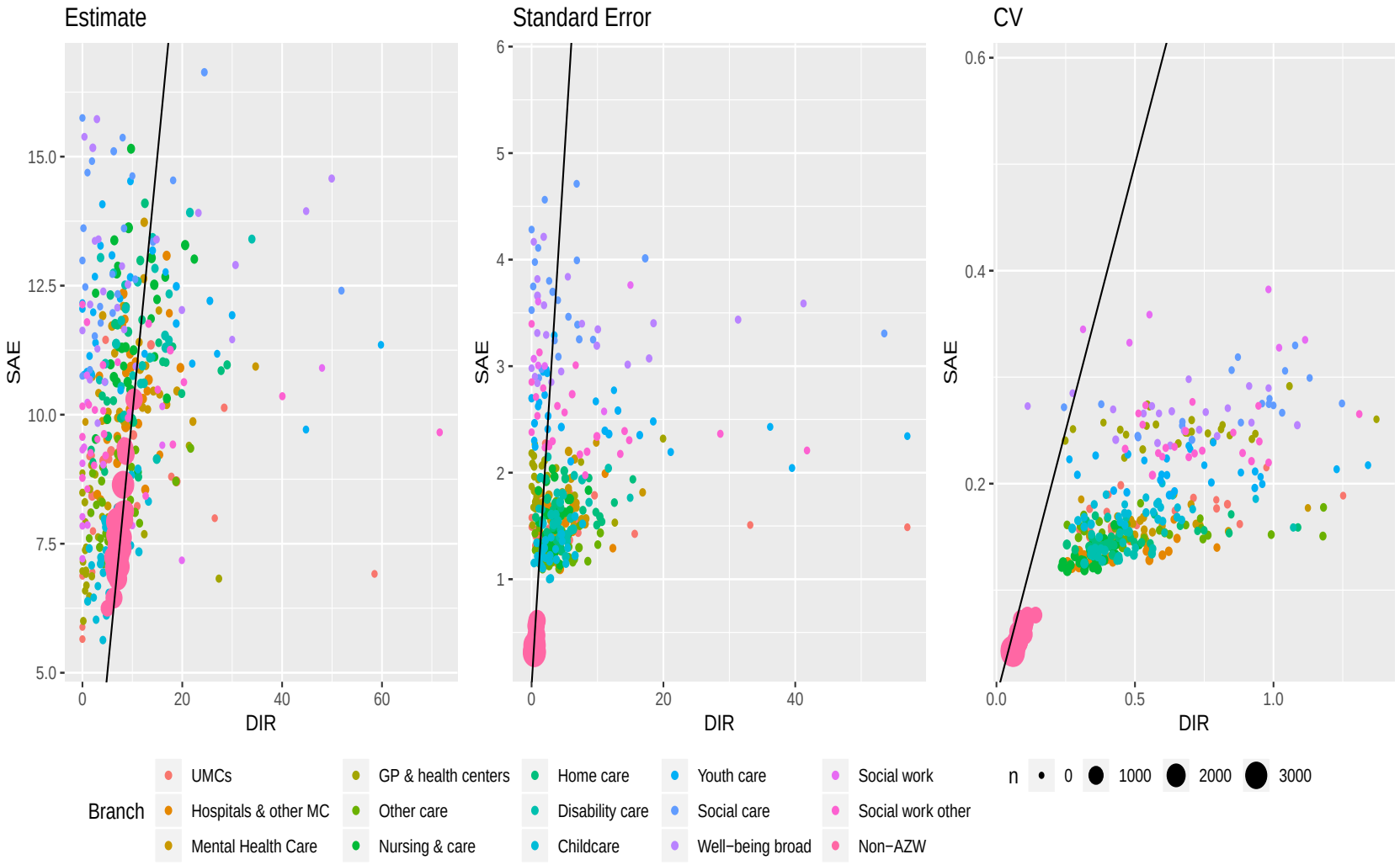

Figure 4.18 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the number of absence days due to sickness at the target small domains cross-classified by region and subbranch.

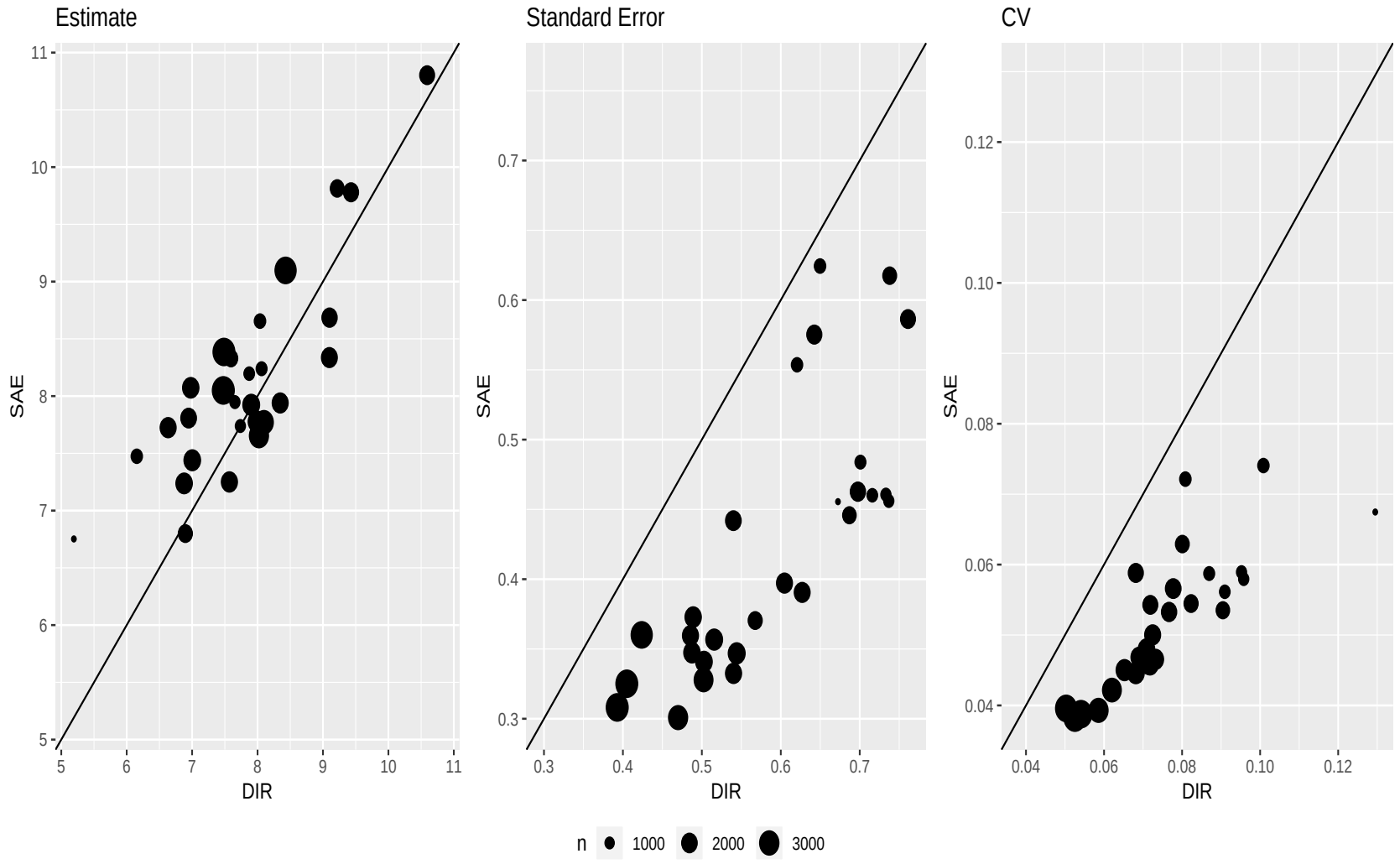

Figure 4.19 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the number of absence days due to sickness at the region level.

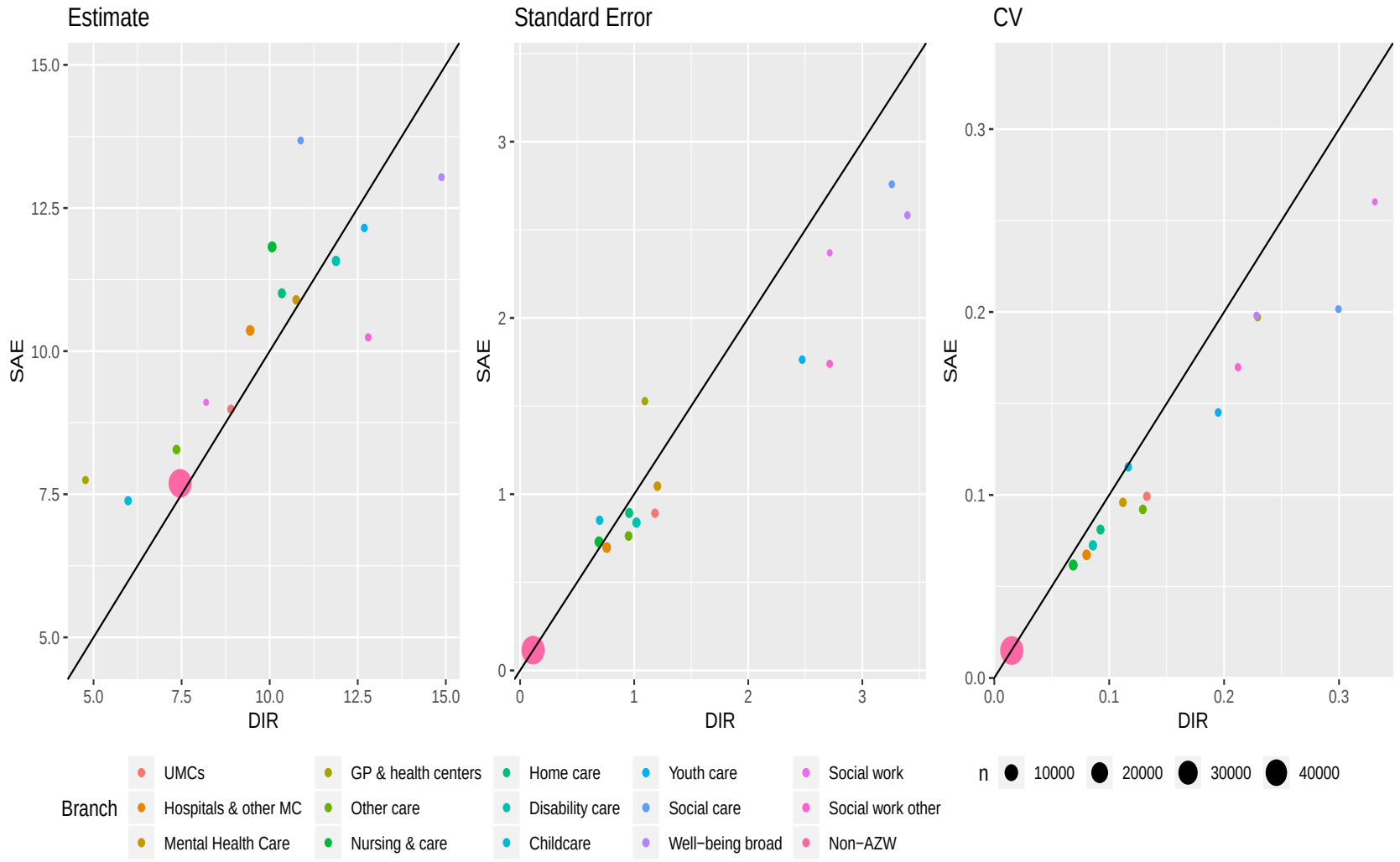

Figure 4.20 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with their standard errors (SE) and coefficient of variation (CV%) for the number of absence days due to sickness at the subbranch level.

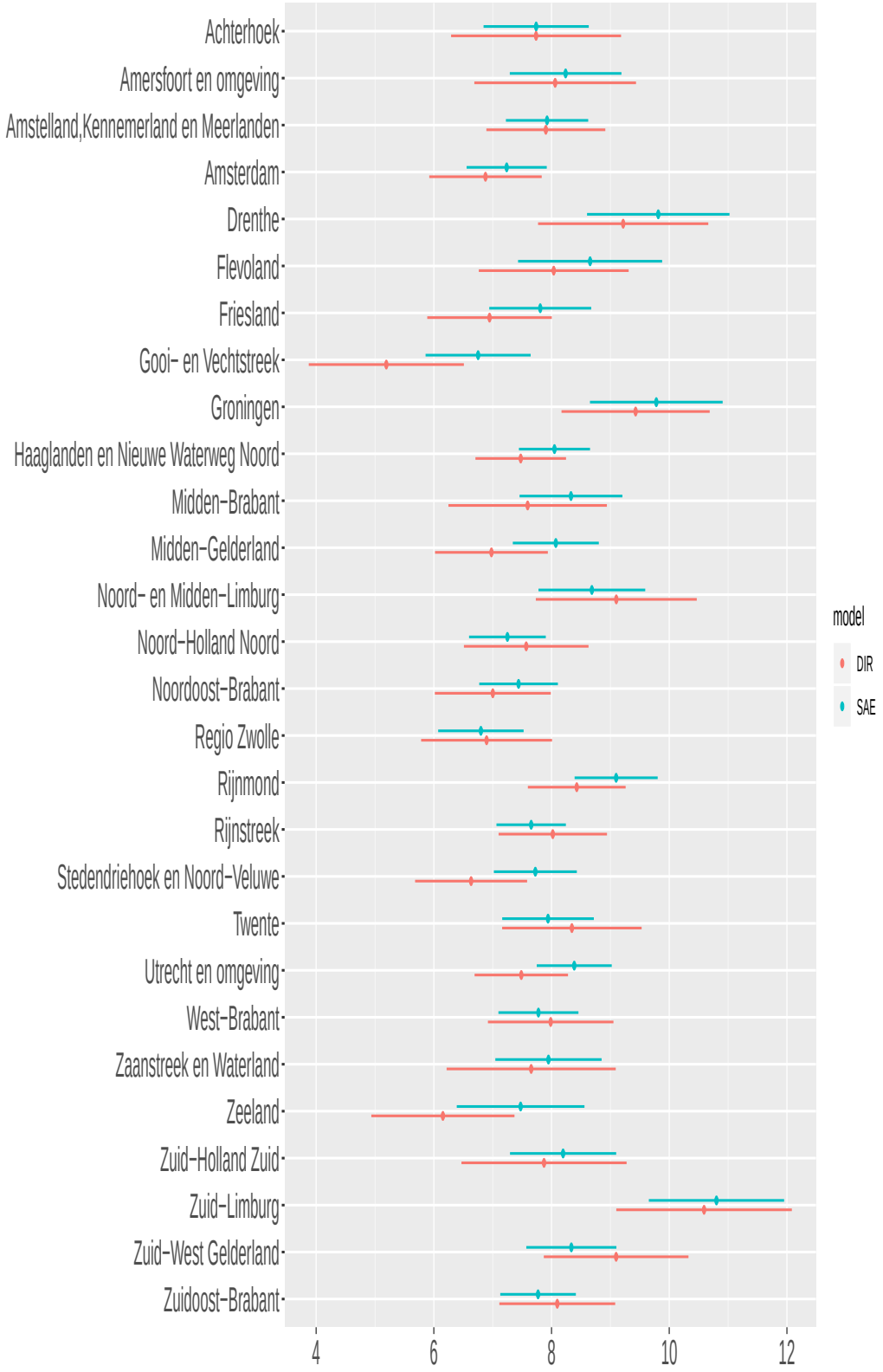

Figure 4.21 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with approximate 95% interval for the number of absence days due to sickness at the region level.

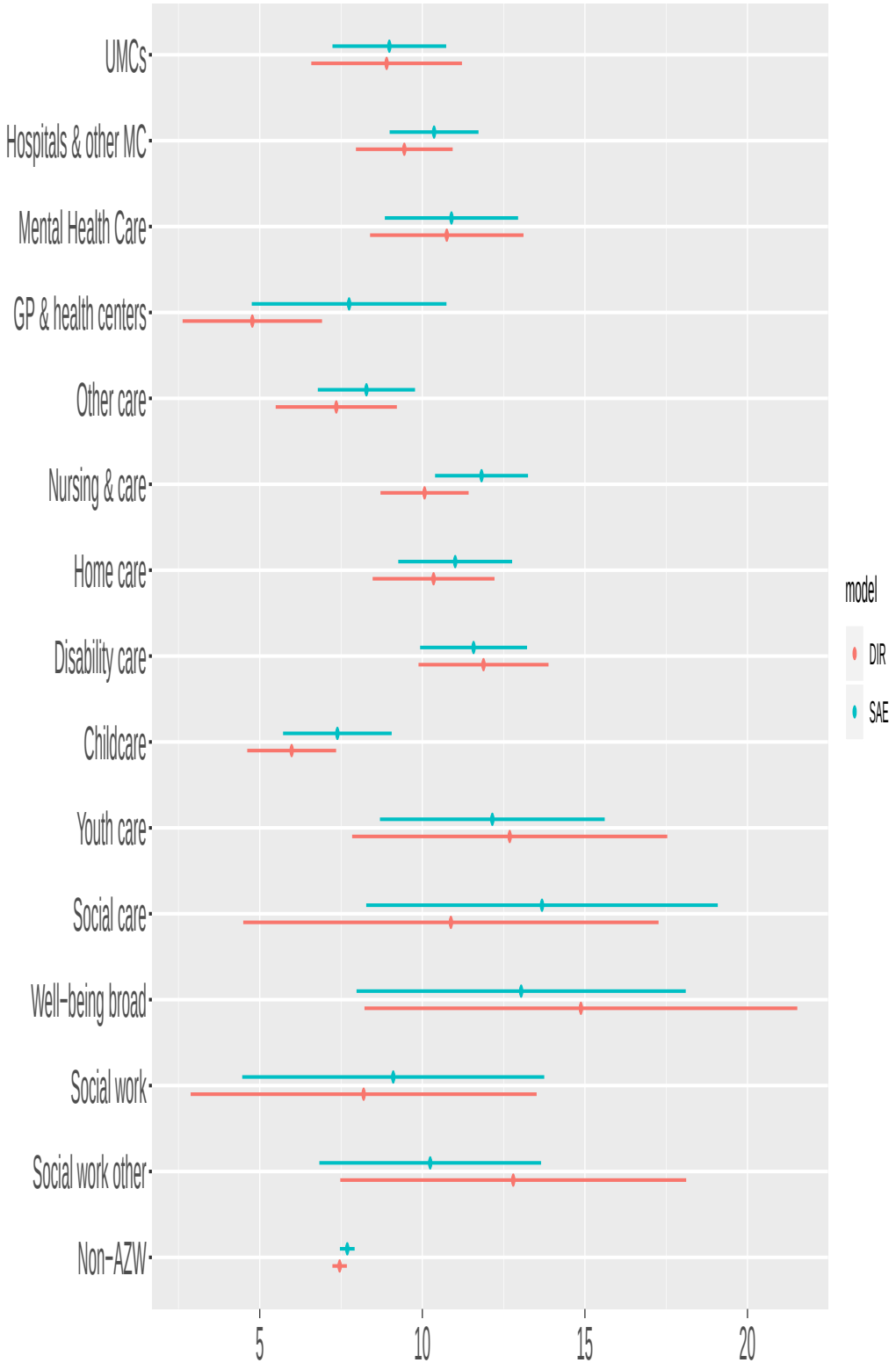

Figure 4.22 Direct (DIR) and model-based (SAE) estimates (negative-binomial multilevel model) with approximate 95% interval for the number of absence days due to sickness at the subbranch level.

4.5 Benchmarking

The model-based small area estimates can be aggregated to the overall level to obtain estimates for the absence indicators for the complete population of employees. However, these figures are already produced based on the NEA weights and published on StatLine. NEA observes many more variables related to working conditions than just the sickness absence variables, and the use of a single set of weights helps to maintain consistency among all derived estimates. The aggregates of the model-based estimates, however, do not exactly agree with the official NEA figures. But the differences are expected to be small since the NEA design and weighting variables are also included in the SAE models used.

Table 4.4 shows the overall official and model-based figures for the four absence indicators considered.

	percentage	binary	frequency	abs. days
NEA (StatLine)	4.48 (0.060)	46.83 (0.207)	1.220 (0.015)	7.84 (0.11)
LIN_S	4.54 (0.065)	46.94 (0.217)	1.231 (0.011)	7.97 (0.11)
LIN_C	4.53 (0.062)	46.64 (0.225)	1.242 (0.010)	7.95 (0.11)
TWOPART_S	4.54 (0.063)			
TWOPART_C	4.54 (0.062)			
BIN_S		46.93 (0.219)		
BIN_C		46.63 (0.219)		
NB_S			1.238 (0.009)	8.07 (0.10)
NB_C			1.256 (0.009)	8.15 (0.12)

Table 4.4 Overall official figures and SAE estimates, with standard errors in parentheses

The differences at the overall level are mostly small. The largest differences between direct (NEA) and model-based estimates can be seen for the count variables (frequency and absence days), where the model-based estimates tend to be higher, especially for the estimates based on the negative binomial model using the complex set of covariates. For the absence days variable this largest relative difference is almost 4%, which is a bit more than expected, and also is not negligible with respect to the standard errors. A partial explanation for these somewhat higher estimates might be that the complex covariate model yields a better correction for selective non-response bias. This is confirmed by the results of a logistic regression using the simple and complex covariate models applied to the NEA response indicator. The logistic regression models fit to the original NEA sample (of size 172615, of which 58316 employees responded) show a clear difference in model fit: the complex covariate model has DIC/WAIC values about 2000 units lower than the simple covariate model, which is very strong evidence for a better explanation of response.

The model-based small area estimates can easily be adjusted so as to be consistent with the official NEA figures at the overall level. This procedure is common in small area estimation and is called benchmarking. We use a procedure that minimizes a weighted sum of squared differences between original and benchmarked small area estimates subject to the single constraint that the aggregate estimate equals the official NEA figure. More precisely, if $\hat{\theta}$ denotes the $(M = 420)$-dimensional vector of small area mean estimates at the RegioPlus × subbranch level, and $\hat{\bar{Y}}$ denotes the official NEA population mean estimate, then the vector of benchmarked small area estimates $\hat{\theta}^{(b)}$ is

found by minimizing

$$(\hat{\theta}^{(b)} - \hat{\theta})' V^{-1} (\hat{\theta}^{(b)} - \hat{\theta}) \tag{17}$$

subject to $R'\hat{\theta}^{(b)} = \hat{\hat{Y}}$, where R is the $M \times 1$ matrix with relative domain sizes N_d/N, and V is the (MCMC estimate of the) posterior covariance matrix for the small area means. The solution is

$$\hat{\theta}^{(b)} = \hat{\theta} + VR(R'VR)^{-1}(\hat{\hat{Y}} - R'\hat{\theta}) . \tag{18}$$

The benchmarked estimates at the RegioPlus or subbranch levels are simply obtained by aggregating $\hat{\theta}^{(b)}$.

Figure 4.23 shows the small (downward) adjustments due to benchmarking for the percentage absence time variable. A similar plot is provided in figure 4.24 for the number of absence days estimated with the negative-binomial model. Here the downward adjustment due to benchmarking is even larger.

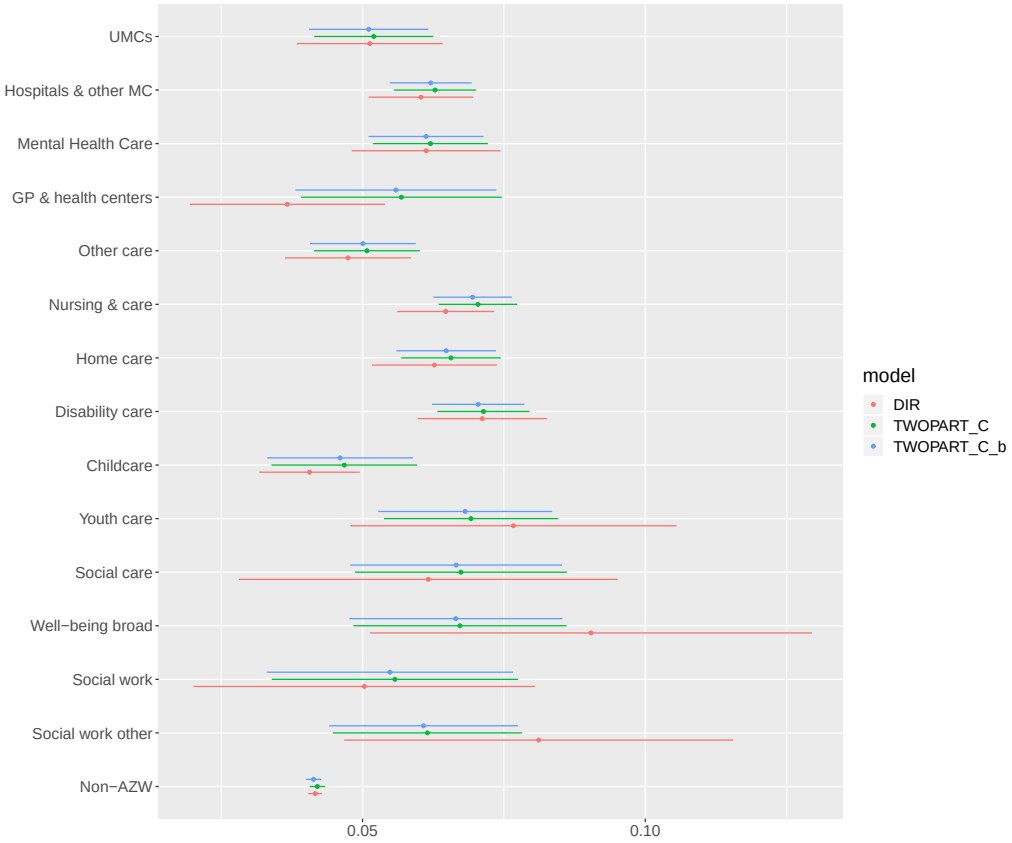

Figure 4.23 Direct (DIR) estimates and both original and benchmarked (two-part multilevel model with complex fixed effects) estimates based on the two-part multilevel model with complex covariate model, for binary absence at the subbranch level, with approximate 95% intervals.

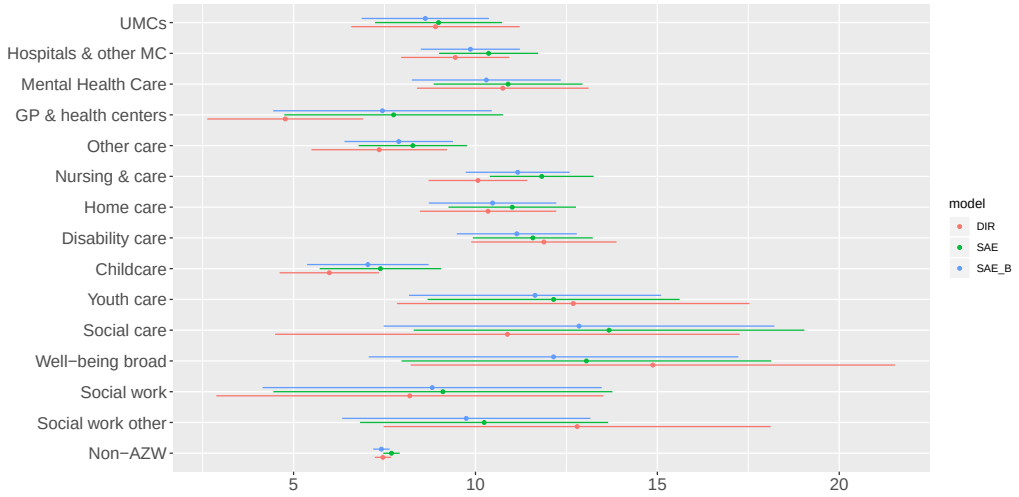

Figure 4.24 Direct (DIR) estimates and both original and benchmarked (negative-binomial multilevel model with complex fixed effects) estimates based on the two-part multilevel model with complex covariate model, for number of absence days at the subbranch level, with approximate 95% intervals.

5 Discussion

A feasibility study has been carried out to show how the method of small area estimation can be used to improve estimates of sickness absence at a detailed level. At the most detailed level of RegioPlus × AZW subbranch there is a substantial improvement over direct estimates that can be obtained using NEA survey weights. However, the relative standard errors at this detailed level can still be rather high. Especially for the count variables (frequency of absence periods and number of absent days) many estimates have coefficients of variation (CVs) higher than 20%. For the binary absence variable, however, all CVs are below 10%. At the RegioPlus and AZW subbranch levels the small area estimates are also more precise than the direct estimates, although the relative gain is smaller.

Several ways to further improve the precision of the small area estimates could be examined in a follow-up study. The most promising way of further improving the estimates is to extend the modeling over several years, thus borrowing strength not only over cross-sectional domains but also over multiple years of NEA data. Note that this method, provided the time-series model components are sufficiently flexible, is quite different from using three-year moving averages, such as is current practice for figures at the RegioPlus and subbranch levels. Not only is strength also borrowed across domains, but the estimates also remain more specific to each year (instead of to the last three years).

Another way to potentially further improve the small area estimates, at least for the percentage absence time, is to also use data from the KVZ quarterly business survey. Based on KVZ, qualitatively good estimates can be obtained by subbranch, although the measured concept is slightly different from that of NEA. One possible way to use KVZ data is to include KVZ estimates at subbranch level as additional covariates in the model, subject to measurement error, by extending the model as described in (Ybarra and Lohr,

2008; Arima et al., 2015).

Finally it appears that the variables are correlated with each other. Therefore, a third way to improve the precision is to combine the different variables in one model and model the correlations between the different random effects. This results in a multivariate unit level model.

Variables for which the CV is smaller than 10%, the point estimates are sufficiently precise to publish. This applies e.g. to the variable percentage of employees with absence at the most detailed level of subbranch and region (CV < 10%), the regional level (CV < 3%) and the subbranch level (CV < 5%). For variables with CVs that vary between 10% and 20% or even 25%, publication might be considered but in that case it is important to publish also the standard errors to highlight their increased uncertainty. This applies e.g. to the variable percentage of absence time at the most detailed level (CV between 8% and 25%) and at the subbranch level (CV < 15%). Variables with CVs larger than 25% we do not recommend publication yet. This applies e.g. to the variable number of absence days at the most detailed level of the crossing of subbranch and region, where CVs vary between 12% and 40%.

As indicated above there is ample room to further improve the models developed so far. From that point of view it can be argued that it is currently too early to publish even a limited set of variables. If the models are indeed improved in a later stage, it might become necessary to revise earlier published tables. In case publication for a limited set of variables is considered, a publication strategy must be developed to address the following issues:

- Revision strategy: if models eventually are improved, are earlier published figures replaced by these updates?
- An alternative is to qualify the set of variables that are already published as experimental statistics that might be revised in the future
- It is recommended to publish besides point estimates also the standard errors, particularly if CVs are larger than 10%
- For tables with a relatively small number of domains with CVs >25% it is also possible to suppress these highly uncertain values in the tables

We finally note that the above-mentioned CV values of 10% and 25% as criteria for publication are somewhat arbitrary and are only intended as a guideline.

Acknowledgement

We wish to thank John Michiels for helping prepare the data and Marc Smeets and Rianne Verwijs for useful comments on this manuscript.

References

Arima, S., G. Datta, and B. Liseo (2015). Bayesian estimators for small area models when auxiliary information is measured with error. *Scandinavian Journal of Statistics 42*(2), 518–529.

Battese, G., R. Harter, and W. Fuller (1988). An error-components model for prediction of

county crop areas using survey and satellite data. *Journal of the American Statistical Association 83*(401), 28–36.

Boonstra, H. J. (2021). *mcmcsae: Markov Chain Monte Carlo Small Area Estimation*. R package version 0.6.0.

Boonstra, H. J., B. Buelens, and M. Smeets (2007). Estimation of municipal unemployment fractions - a simulation study comparing different small area estimators. BPA nr. DMK-DMH-2007-04-20-HBTA, Statistics Netherlands.

Chandra, H. and U. Sud (2012). Small area estimation for zero-inflated data. *Communications in Statistics - Simulation and Computation 41*, 632–642.

de Vries, J. and J. Michiels (2019). Haalbaarheidsonderzoek kleinedomeinschatters positie in de werkkring en onderwijsdeelname. report version 17-12-2019, in Dutch, Statistics Netherlands.

Gelfand, A. and A. Smith (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*(3), 515–533.

Gelman, A., J. Carlin, H. Stern, D. B. Dunson, A. Vehtari, and D. Rubin (2004). *Bayesian Data Analysis*. Chapman and Hall, New York.

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell. 6*, 721–741.

Hooftman, W., G. Mars, J. Knops, L. van Dam, E. de Vroome, B. Janssen, A. Pleijers, and S. van den Bossche (2020). Nationale enquête arbeidsomstandigheden 2019. Methodological report, in Dutch, TNO and CBS.

Krieg, S., H. Boonstra, and M. Smeets (2016). Small-area estimation with zero-inflated data - a simulation study. *Journal of Official Statistics 32*(4), 963–986.

Michiels, J. (2014). Consistentie cijfers over ziekteverzuim. internal note, in Dutch, Statistics Netherlands.

O'Malley, A. and A. Zaslavsky (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association 103*(484), 1405–1418.

Pfeffermann, D., B. Terryn, and F. Moura (2008). Small area estimation under a two part random effects model with application to estimation of literacy in developing countries. *Survey Methodology 34*, 67–72.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association 108*(504), 1339–1349.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, J. and I. Molina (2015). *Small Area Estimation*. Wiley-Interscience.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B 64*(4), 583–639.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11*, 3571–3594.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research 14*, 867–897.

Ybarra, L. and S. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika 95*(4), 919–931.

Zhou, M. and L. Carin (2015). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 307–320.

Zhou, M., L. Li, D. Dunson, and L. Carin (2012). Lognormal and gamma mixed negative binomial regression. In *Machine learning: proceedings of the International Conference. International Conference on Machine Learning*, Volume 2012, pp. 1343. NIH Public Access.

Appendix A Covariates used in the multilevel models

Variable	Description	Categories
sex	Sex	male , female
ageclass	Age class (years)	15-24, 25-34, 35-44, 45-54, 55-64, 65-74
ethn	Ethnicity	Native, Western - 1st gen., Western-2nd NonWestern-1st, NonWestern-2nd
Prov	Province	12 provinces
Stratum	Strata of sampling design	41 strata[3)]
Urban	Urbanization	very urban, highly urban, moderately urban, little urban, non-urban
Edu	Education	Primary education, VMBO-b/k MBO1, VMBO-G/T AVO lower secondary education, MBO2 and MBO3, MBO4, Havo - VWO, HBO-WO-bachelor, HBO- WO-master and doctor, Unknown
AZW	Type of health sector	AZW, non-AZW
nonwestern	Western?	Yes, No
typehh	Household type	Single household, Unmarried couple without children, Married couple without children, Unmarried couple with children,Married couple with children, Single parent household, Unknown
income	Income level	(-Inf,1e+03], (1e+03,2e+03],(2e+03,3e+03], (3e+03,5e+03],(5e+03, Inf], Unknown
rate	wage rate per hour	(-Inf,10], (10,15], (15,20], (20,25], (25,40], (40, Inf], Unknown
nmedclass	Number of medication class	0, 1, 2, 3, 4-5, 6-8, 9+
SEC	Socioeconomic status	Employee, Independent, Unemployment benefit + other, Social assistance benefit + incapacity for work, Pension benefit, Other", Unknown
contract	Type of contract	Fixed time, Indefinite time, Not applicable, Unknown
overtime	Doing overtime work?	Yes, No, Unknown
jobtype	Type of job	Head leader of big stockholders, Intern, WSW-er, Temporary worker, On-call worker, Rest, Unknown

Table A.1 Covariates used and their categories

B AZW subbranch names in English and Dutch

1	UMCs	UMCs
2	Hospitals & other MC	Ziekenhuizen en overige medisch specialistische zorg
3	Mental Health Care	Geestelijke gezondheidszorg
4	GP & health centers	Huisartsen en gezondheidscentra
5	Other care	Overige zorg
6	Nursing & care	Verpleging en verzorging
7	Home care	Thuiszorg
8	Disability care	Gehandicaptenzorg
9	Childcare	Kinderopvang (inclusief peuterspeelzaalwerk)
10	Youth care	Jeugdzorg
11	Social care	Maatschappelijke opvang met overnachting
12	Well-being broad	Welzijn breed
13	Social work	Maatschappelijk werk
14	Social work other	Sociaal werk overig
15	Non-AZW	Niet-AZW

C Average absence time

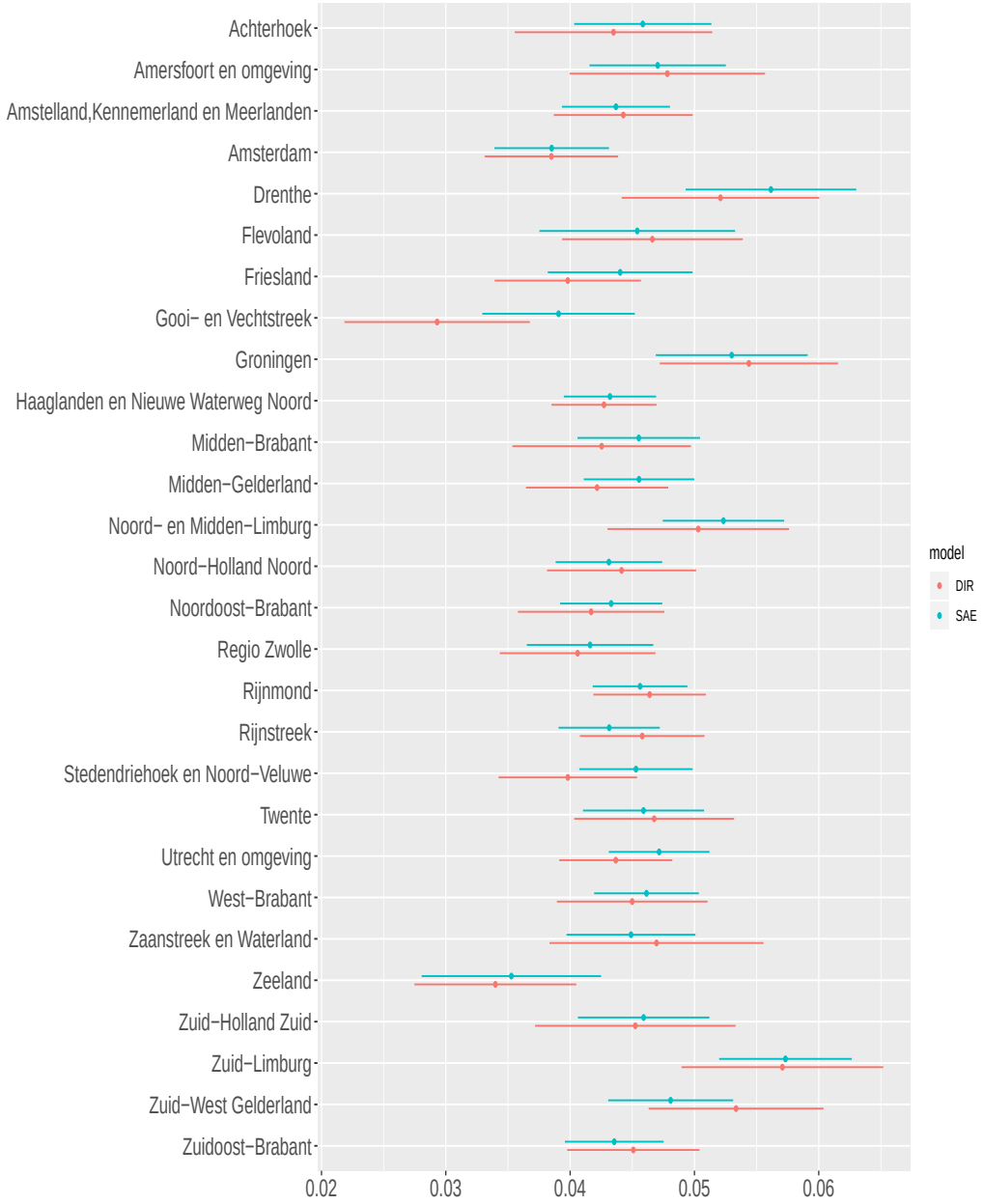

Figure C.1 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage absence time due to sickness at the region level.

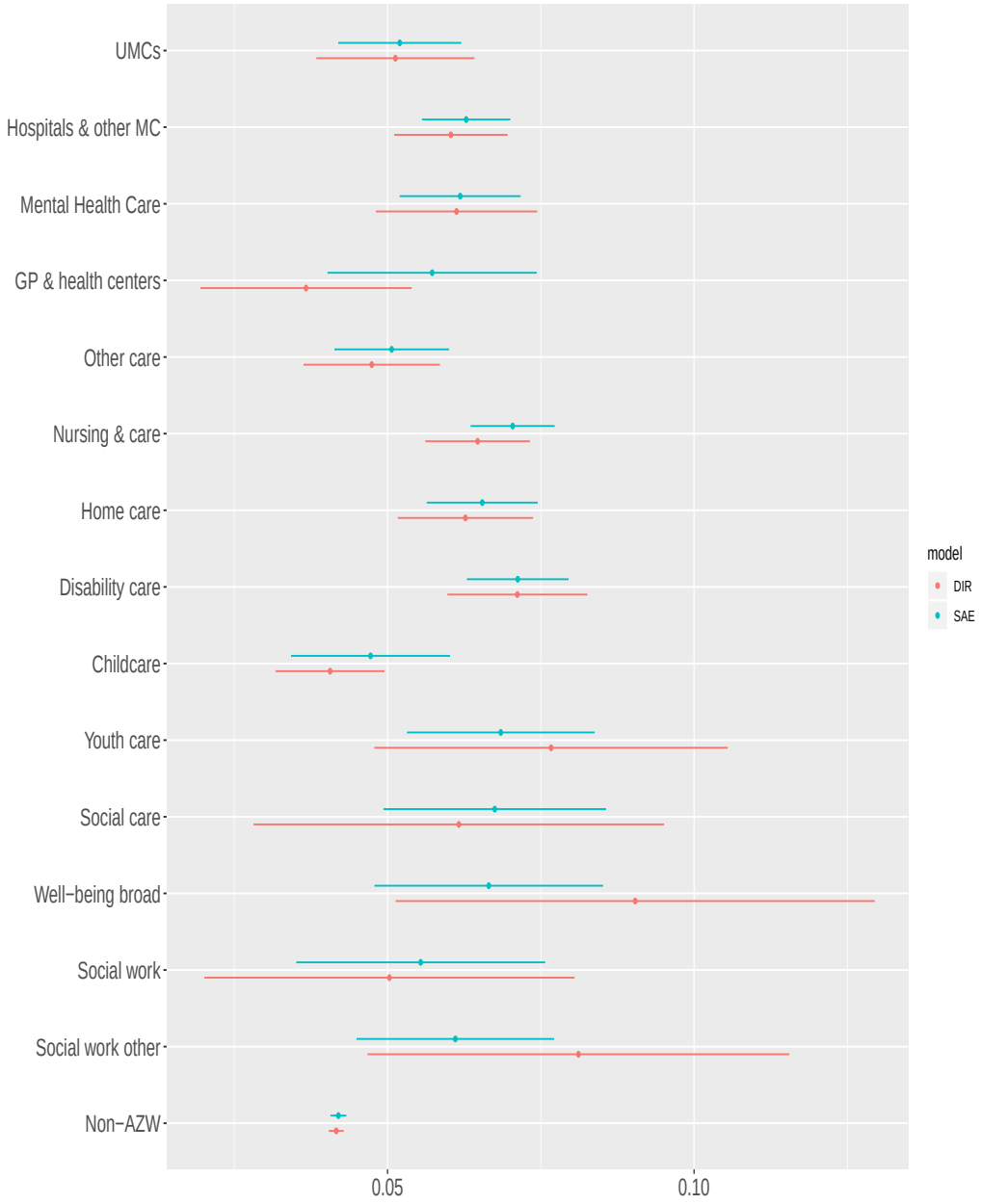

Figure C.2 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage absence time due to sickness at the subbranch level.

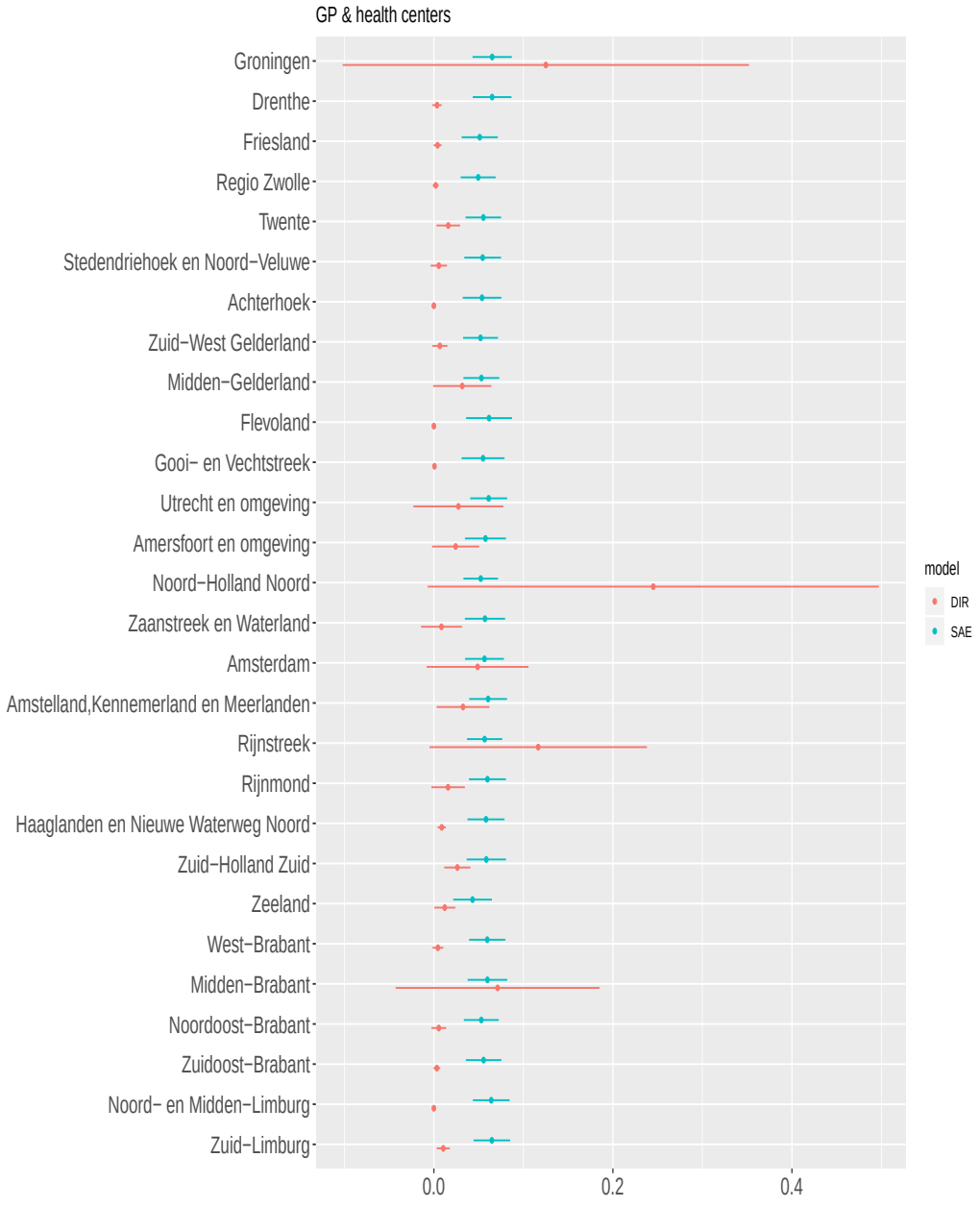

Figure C.3 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage absence time due to sickness at the region level for the "GP and health centers" subbranch.

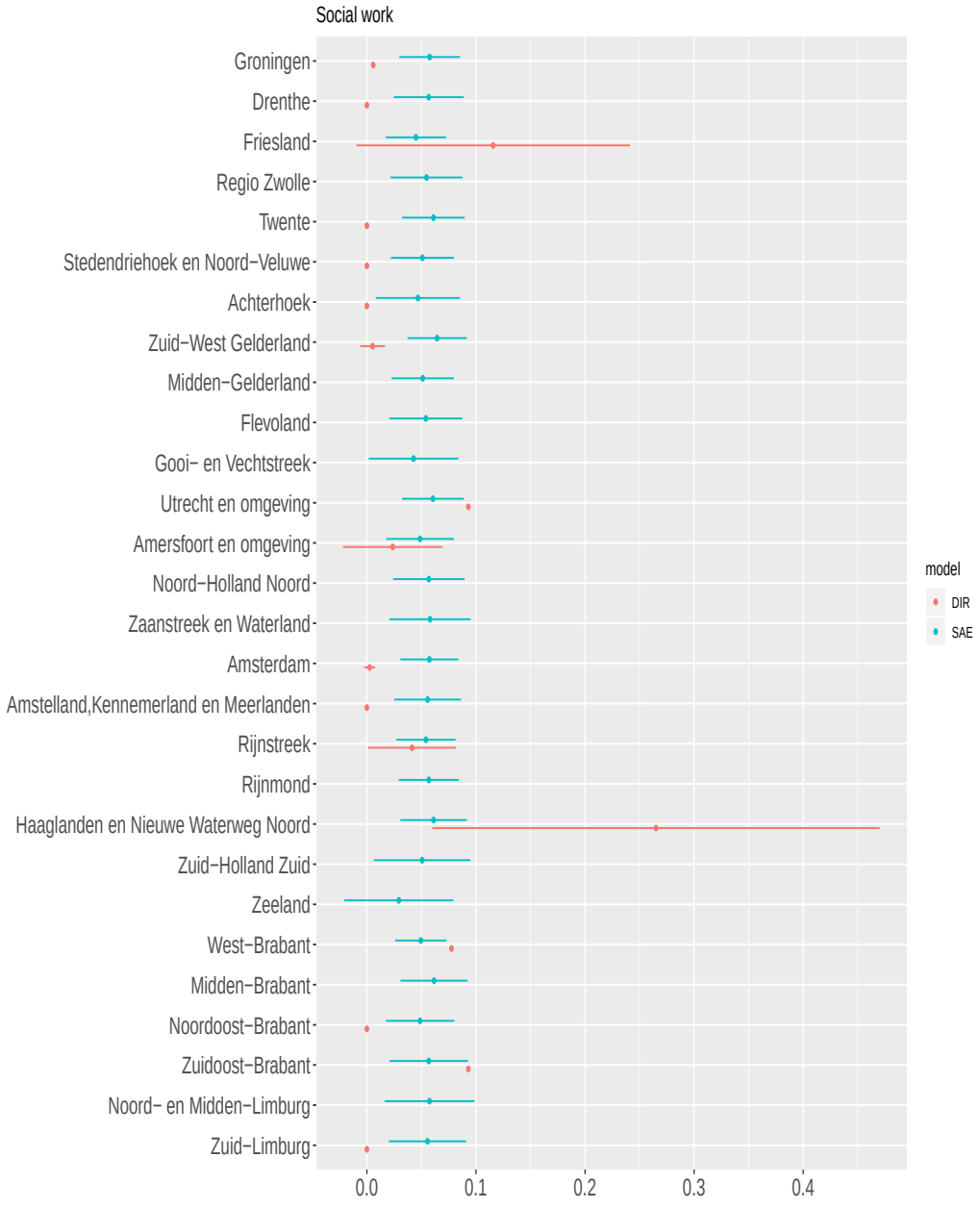

Figure C.4 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage absence time due to sickness at the region level for the "Social Work" subbranch.

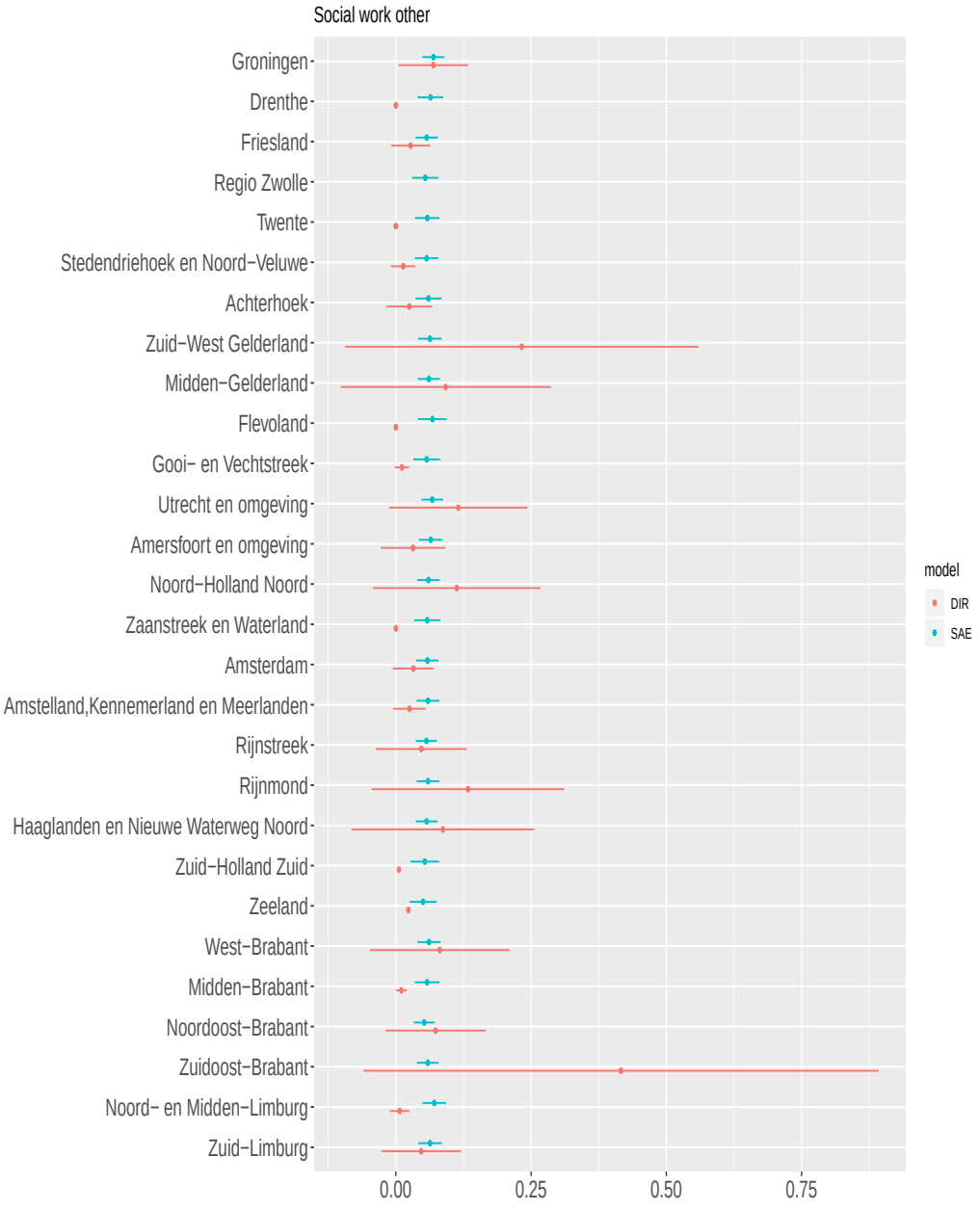

Figure C.5 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage absence time due to sickness at the region level for the "Other Social Work" subbranch.

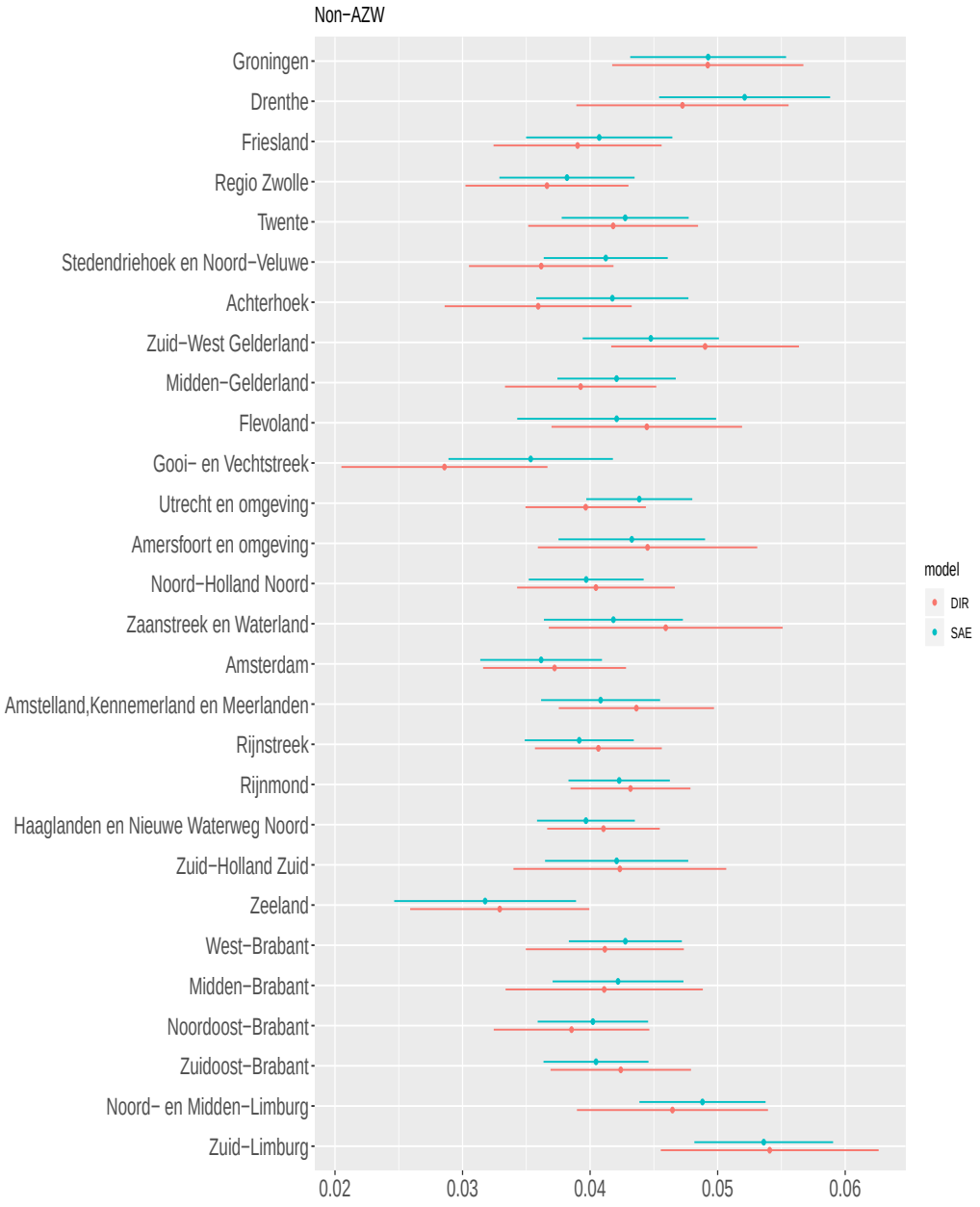

Figure C.6 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage absence time due to sickness at the region level for the "Non-AZW" subbranch.

D Percentage of employees with absence

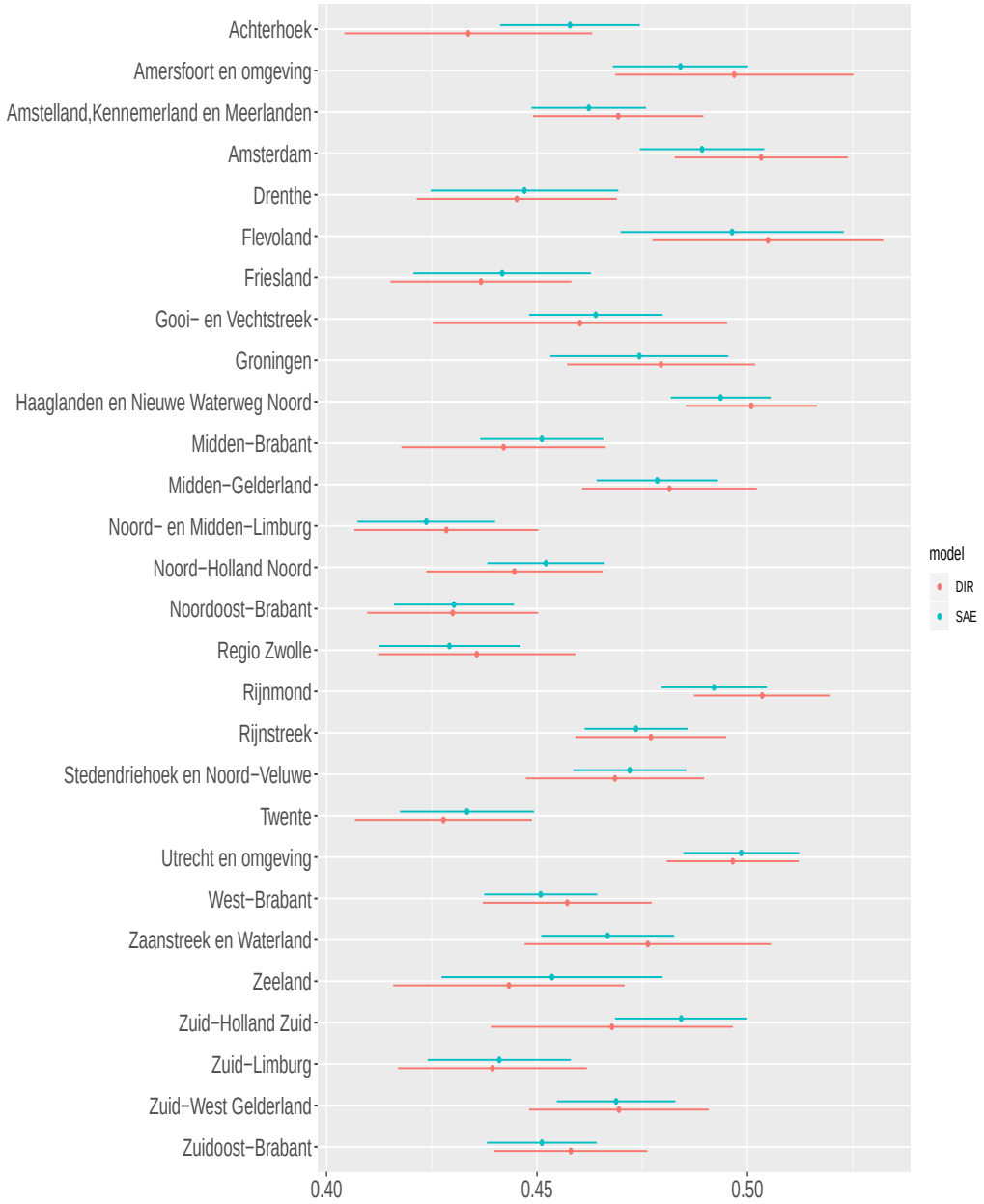

Figure D.1 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage of absent employees due to sickness at the region level.

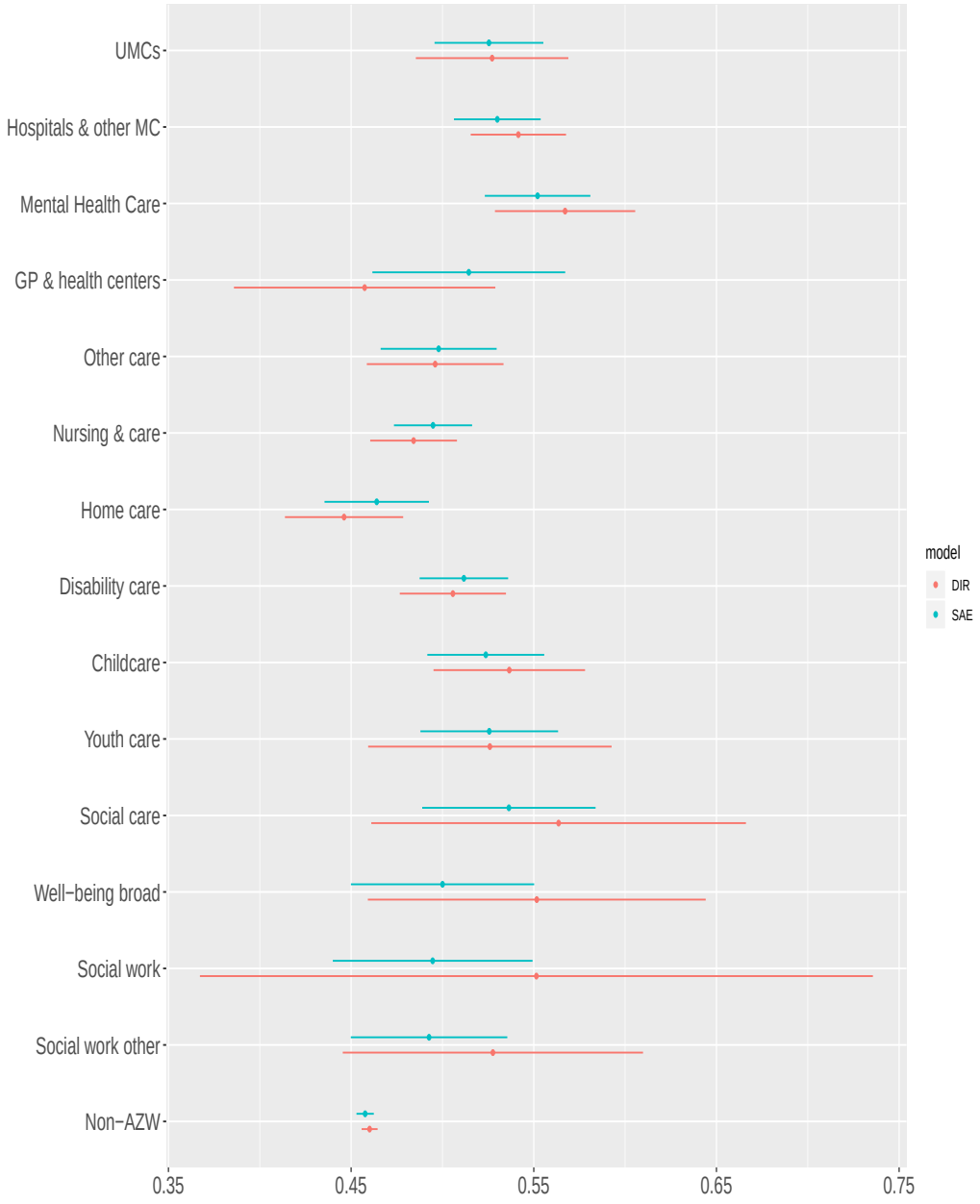

Figure D.2 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage of absent employees due to sickness at the subbranch level.

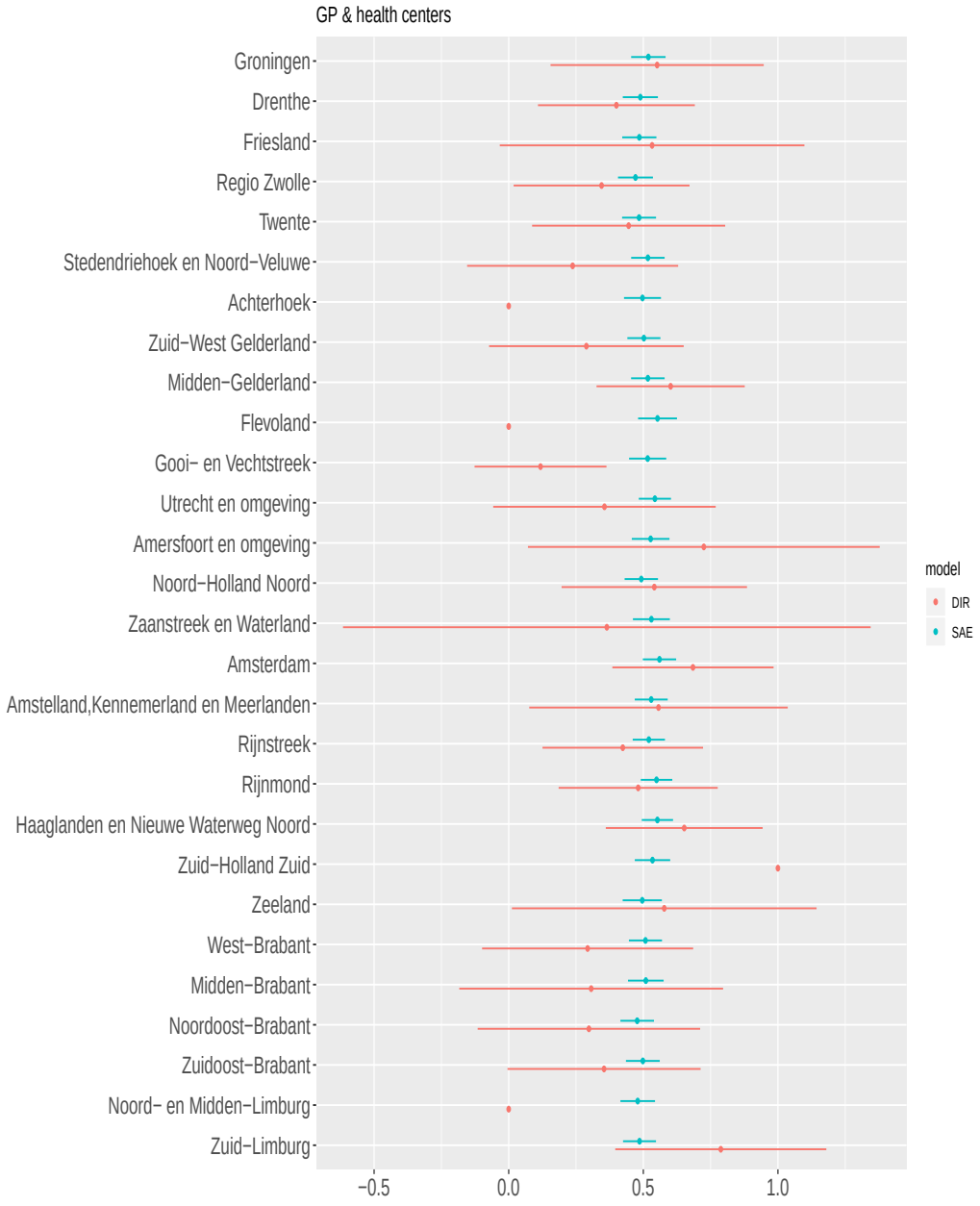

Figure D.3 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage of absent employees due to sickness at the region level for the "GP and health centers" subbranch.

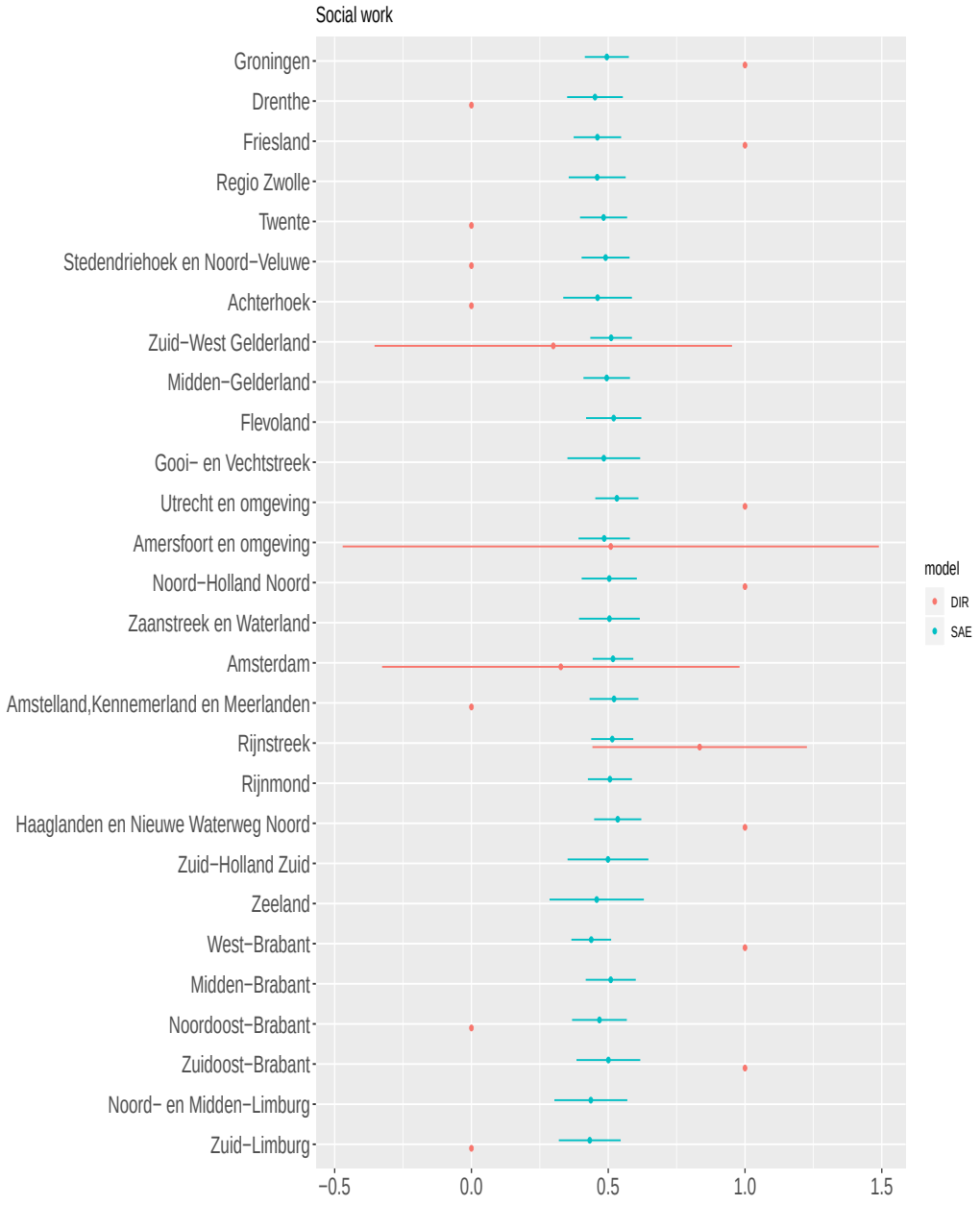

Figure D.4 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage of absent employees due to sickness at the region level for the "Social Work" subbranch.

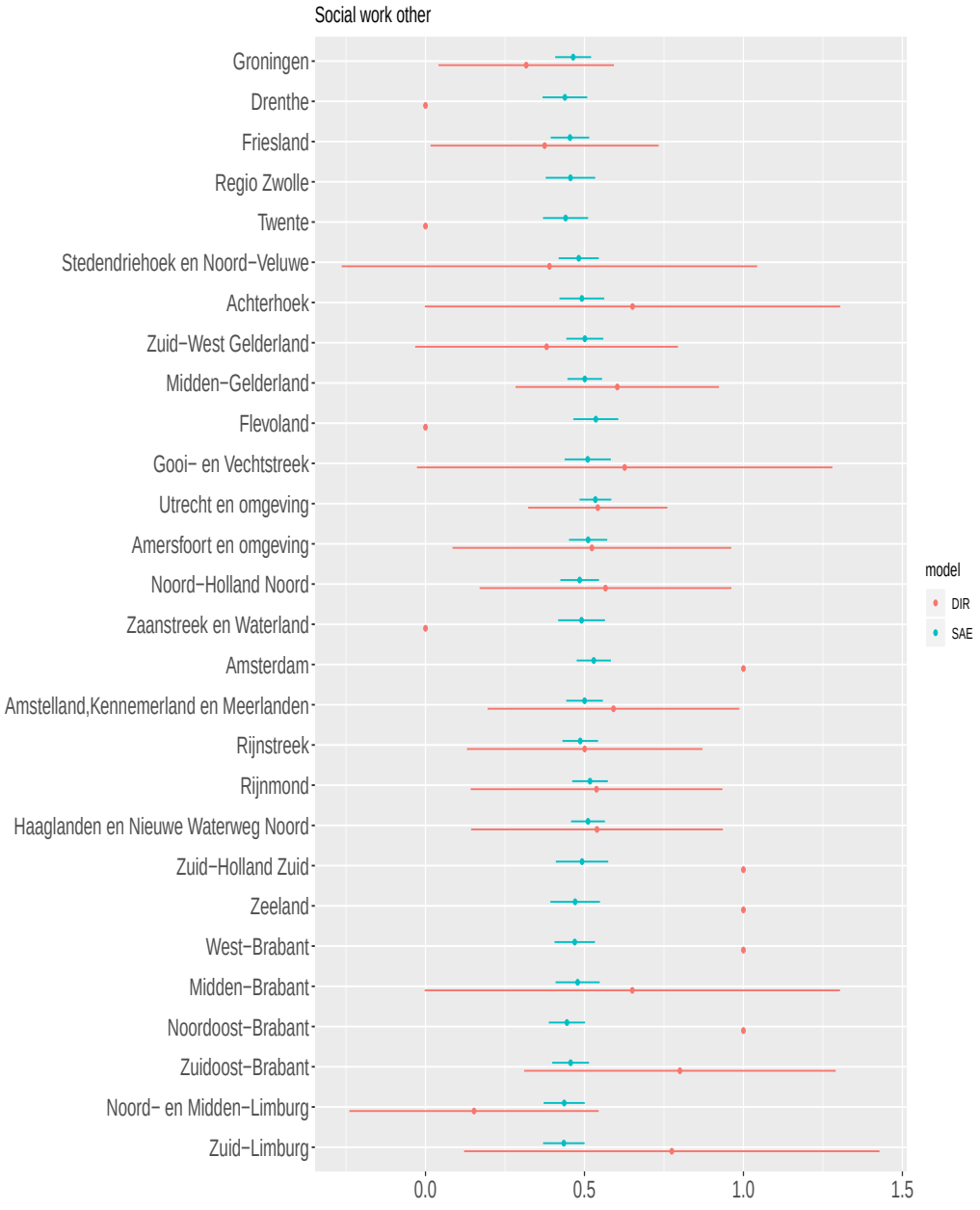

Figure D.5 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage of absent employees due to sickness at the region level for the "Other Social Work" subbranch.

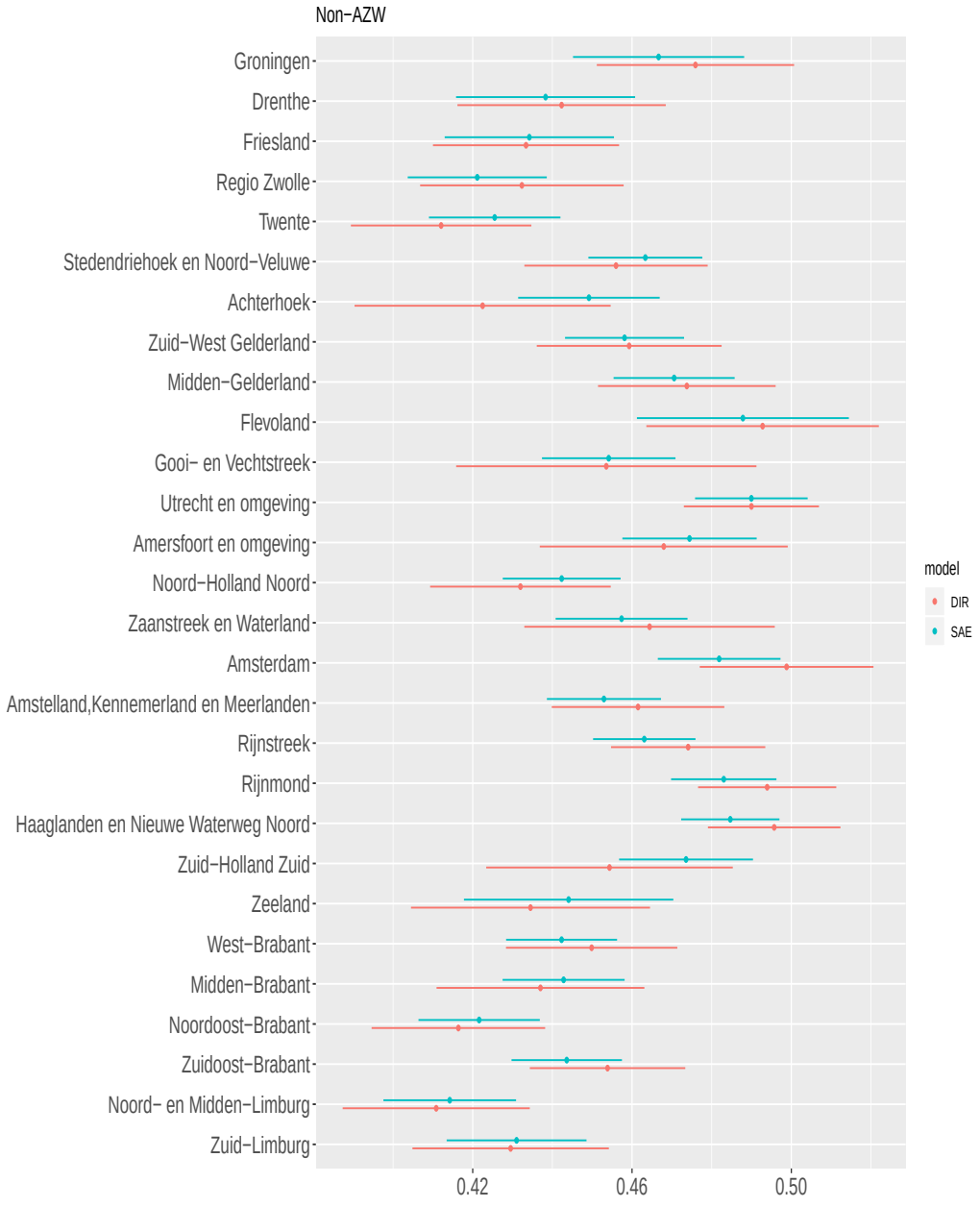

Figure D.6 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the percentage of absent employees due to sickness at the region level for the "Non-AZW" subbranch.

E Number of absent periods

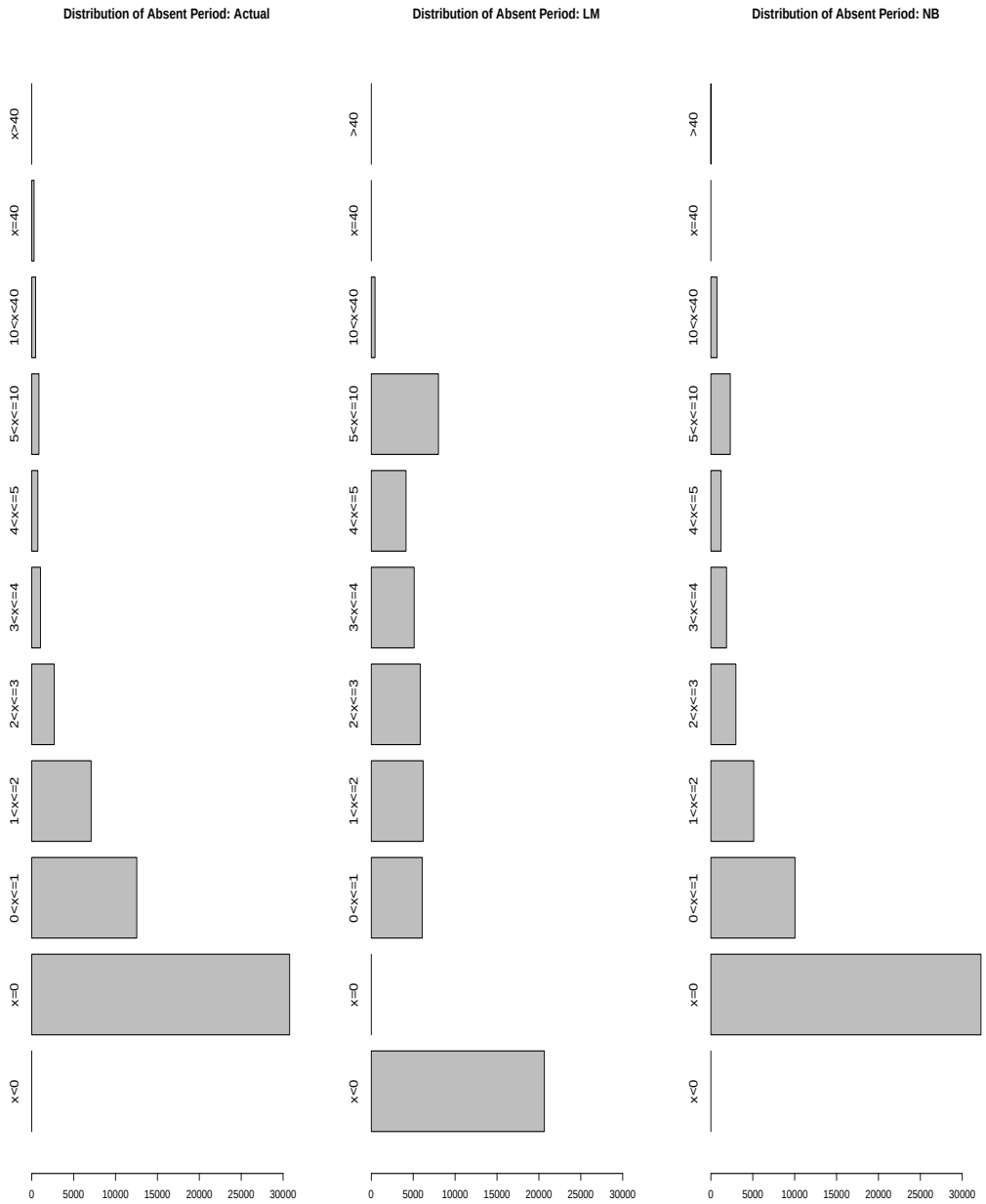

Figure E.1 Distribution of the number of absent periods due to sickness estimated by the linear (LIN) and negative binomial (NB) models with the complex fixed effects component.

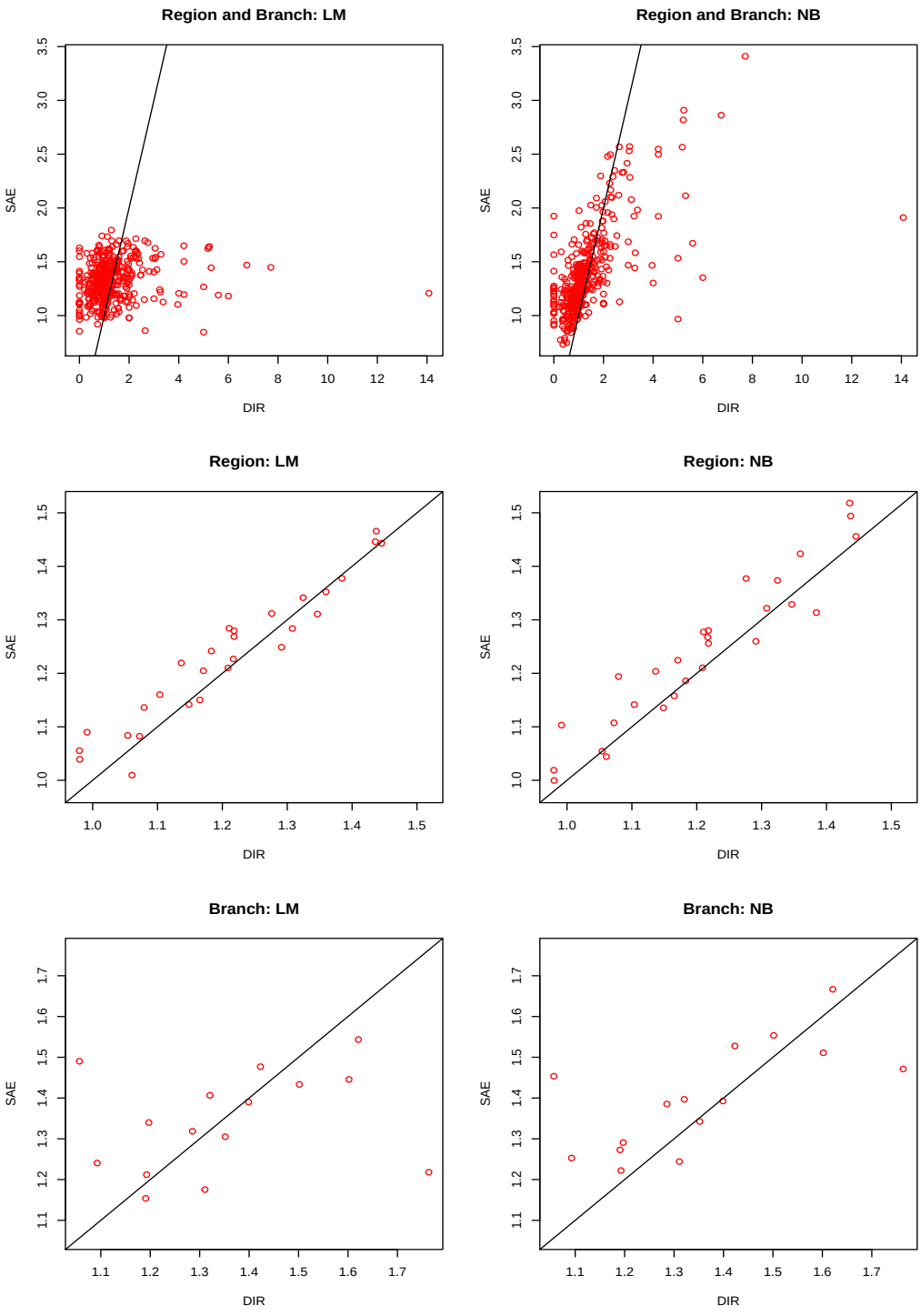

Figure E.2 Estimated mean number of absent periods due to sickness at the region, the subbranch and their cross-classified domains levels based on the linear (LIN) and negative binomial (NB) multilevel models with the complex fixed effects component.

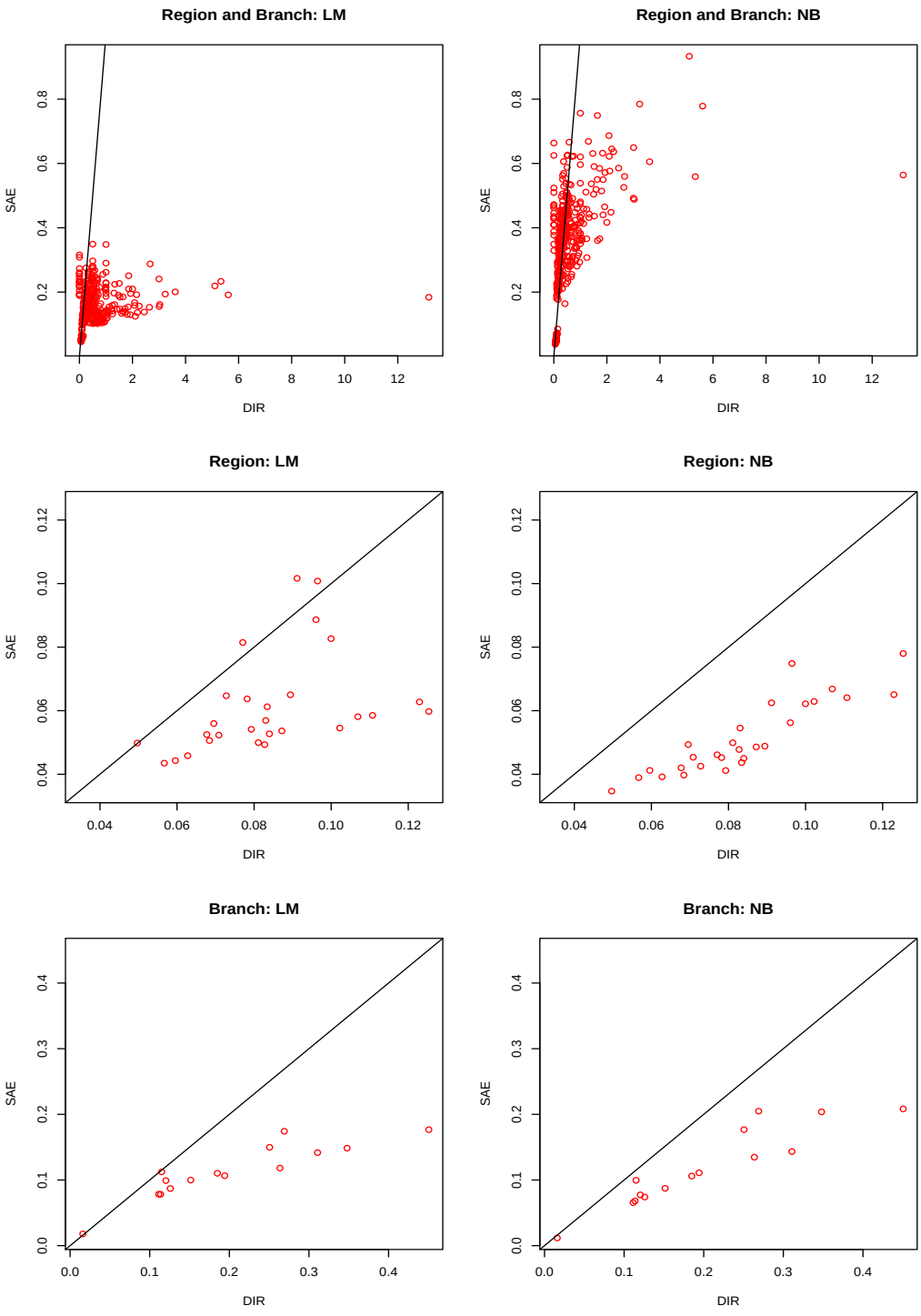

Figure E.3 Estimated standard error (SE) of the estimated mean number of absent periods due to sickness at the region, the subbranch and their cross-classified domains levels based on the linear (LIN) and negative binomial (NB) multilevel models with the complex fixed effects component.

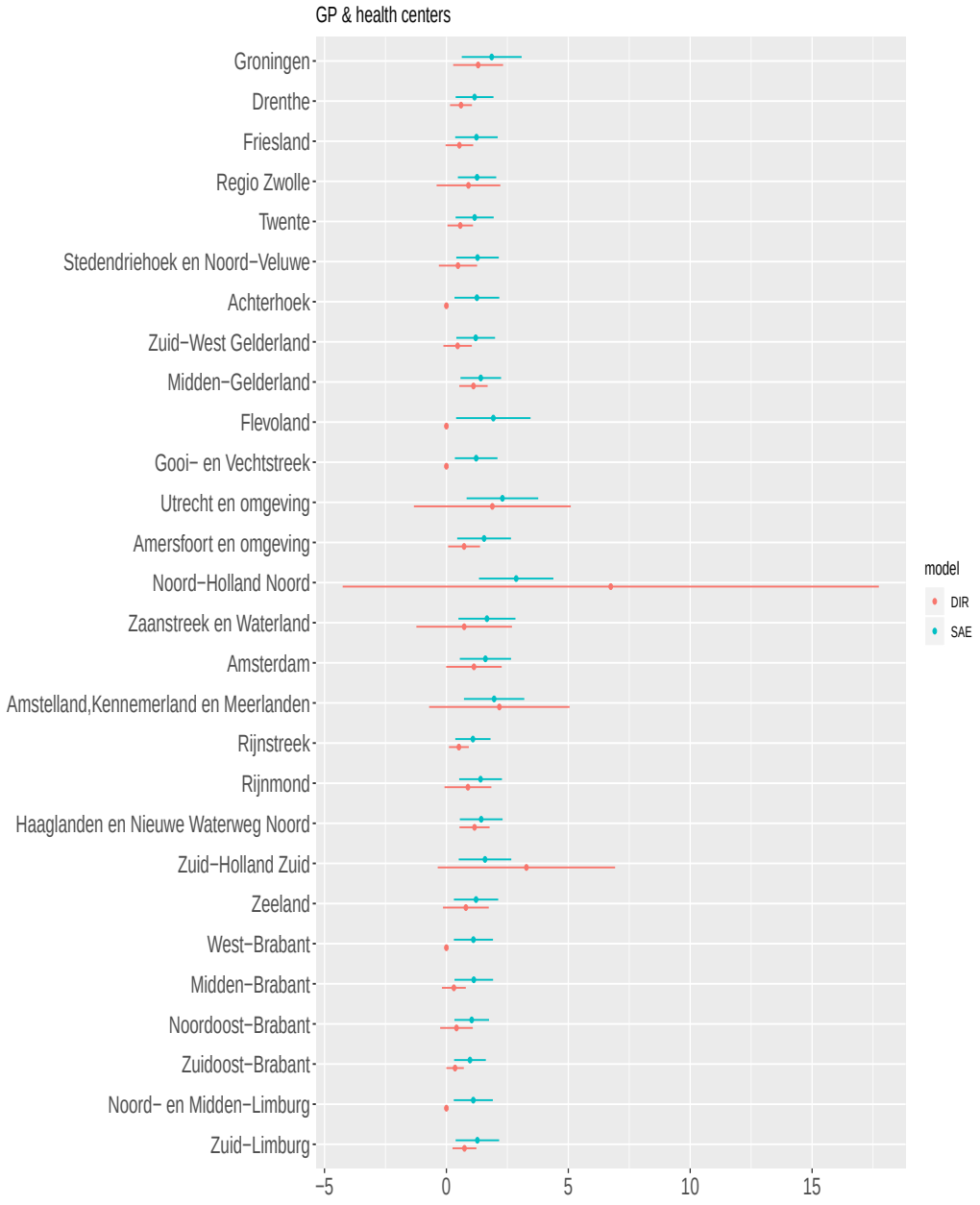

Figure E.4 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the number of absence periods due to sickness at the region level for the "GP and health centers" subbranch.

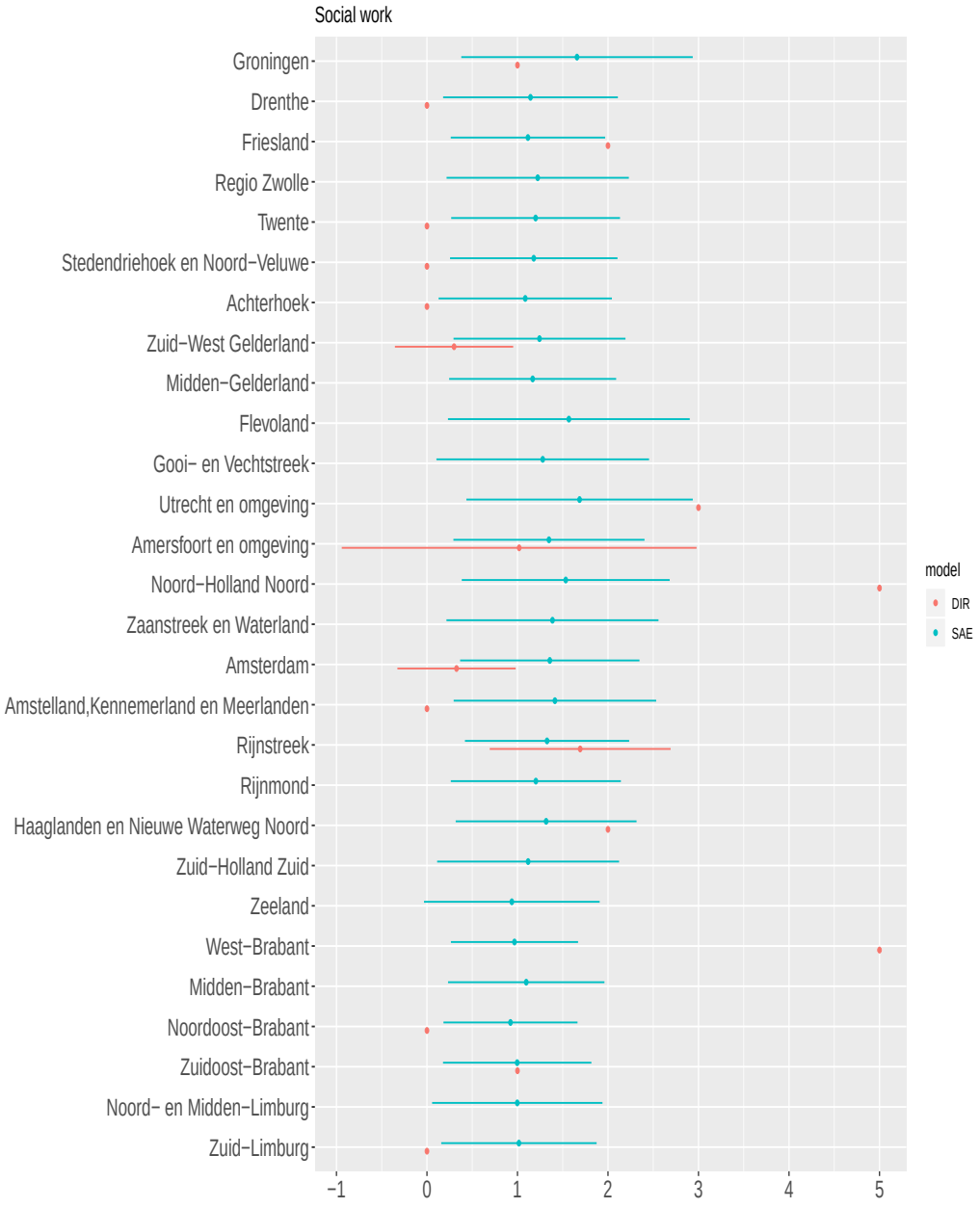

Figure E.5 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the number of absence periods due to sickness at the region level for the "Social Work" subbranch.

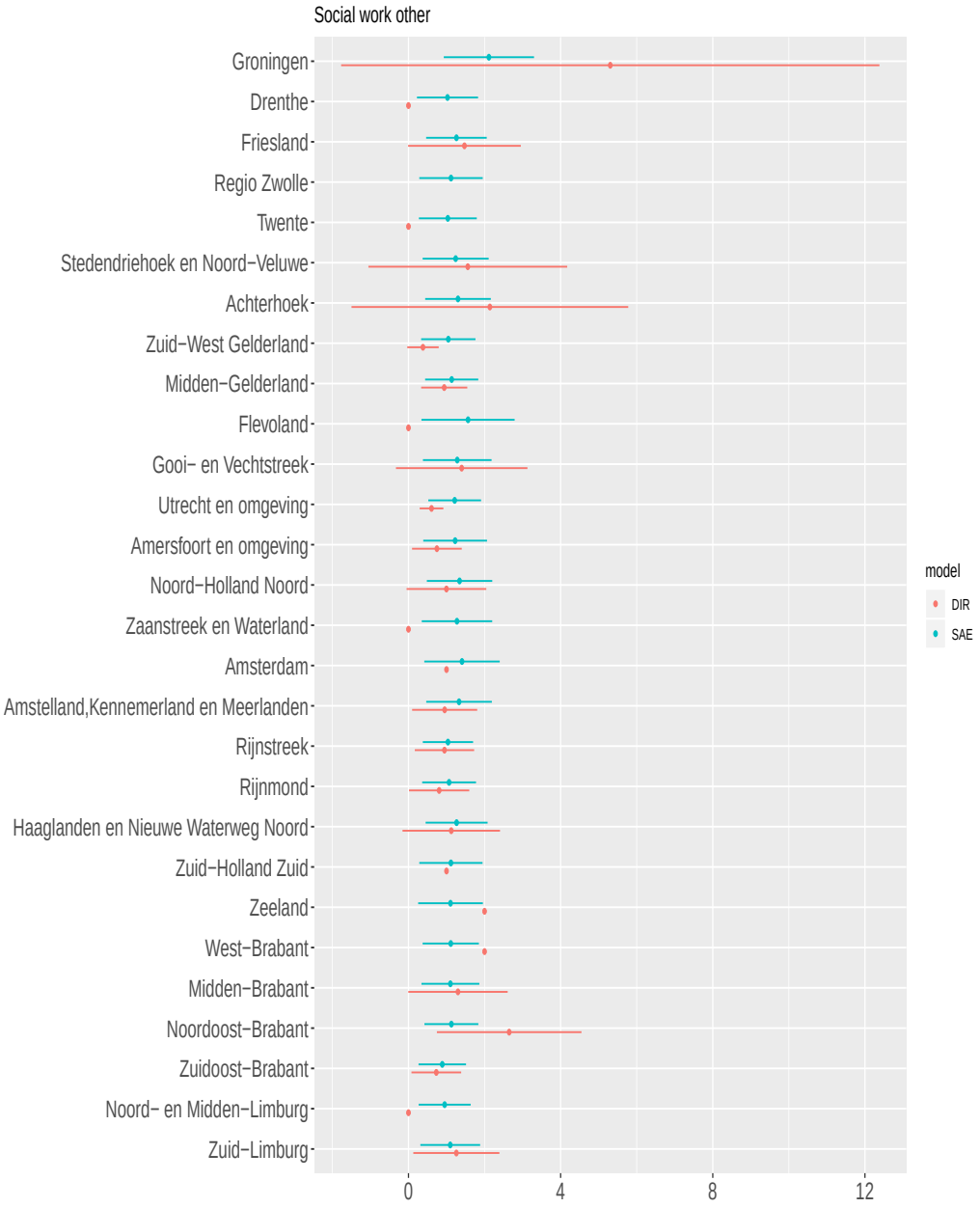

Figure E.6 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the number of absence periods due to sickness at the region level for the "Other Social Work" subbranch.

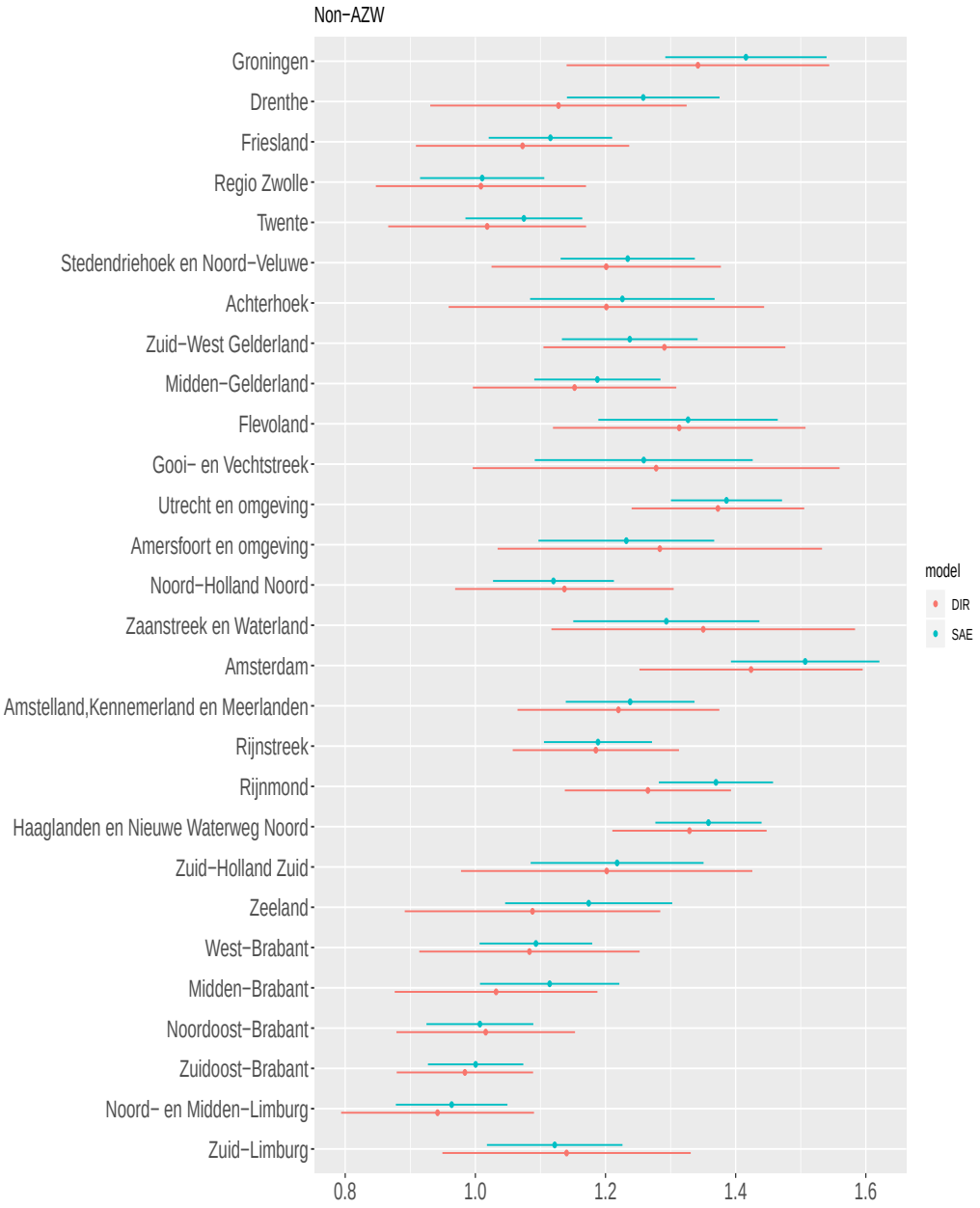

Figure E.7 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the number of absence periods due to sickness at the region level for the "Non-AZW" subbranch.

F Number of absent days

statistics	LIN_S	LIN_C	NB_S	NB_C
DIC	532594	530224	261161	258415
WAIC	532613	530242	261234	258543
σ	25.4	24.9		
r			0.16	0.17
σ_u	0.41	0.28	0.05	0.04
σ_v	2.48	2.38	0.32	0.25
σ_w	0.49	0.43	0.07	0.09

Table F.1 Model information criteria DIC, WAIC and posterior means of standard deviation parameters for linear (LIN) and negative binomial (NB) multilevel models using simple (S) or complex (C) covariate models for the number of absent days

Figure F.1 Distribution of the number of absent days due to sickness estimated by the linear (LM) and negative binomial (NB) models with the complex fixed effects component.

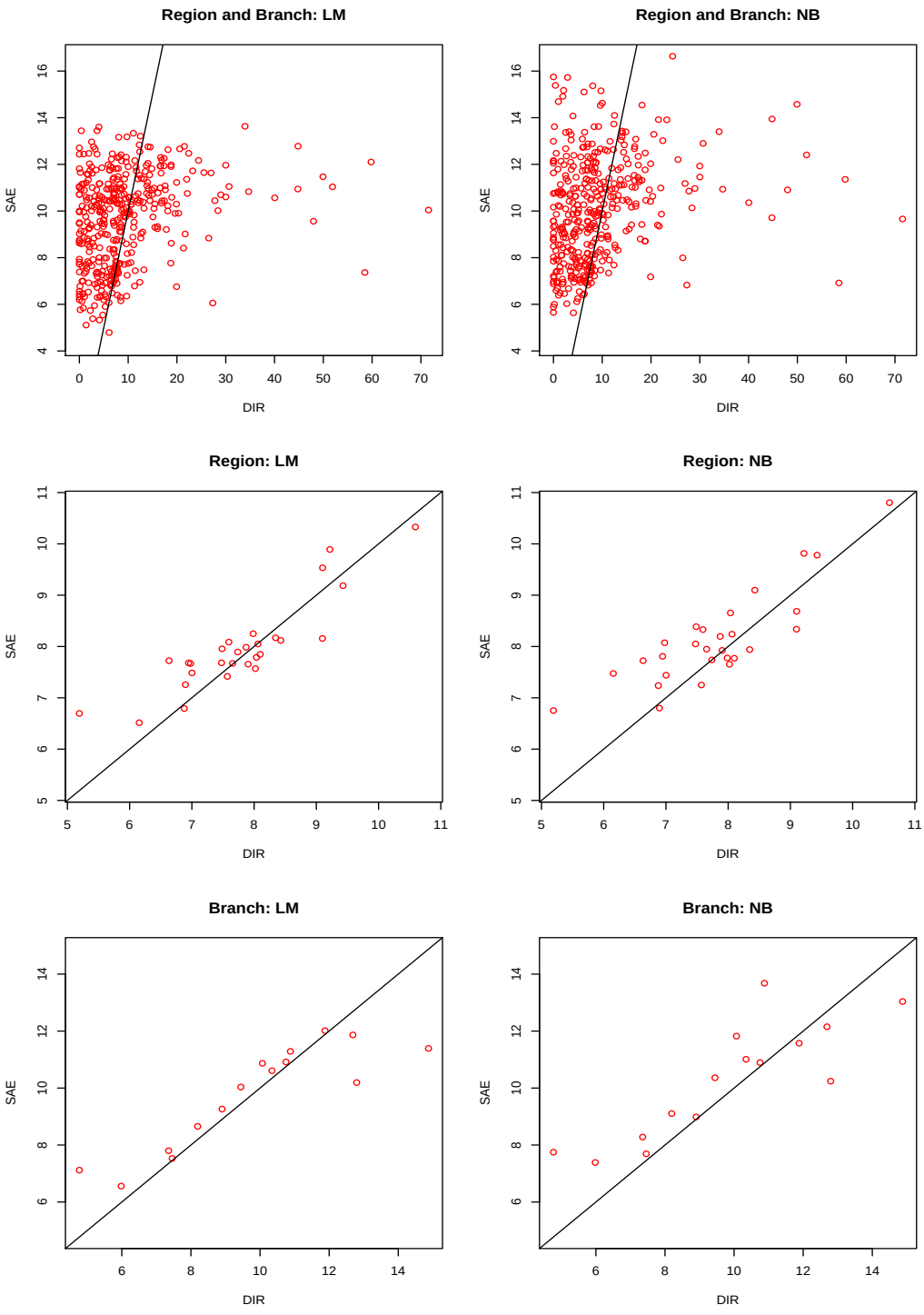

Figure F.2 Estimated mean number of absent days due to sickness at the region, the subbranch and their cross-classified domains levels based on the linear (LIN) and negative binomial (NB) multilevel models with the complex fixed effects component.

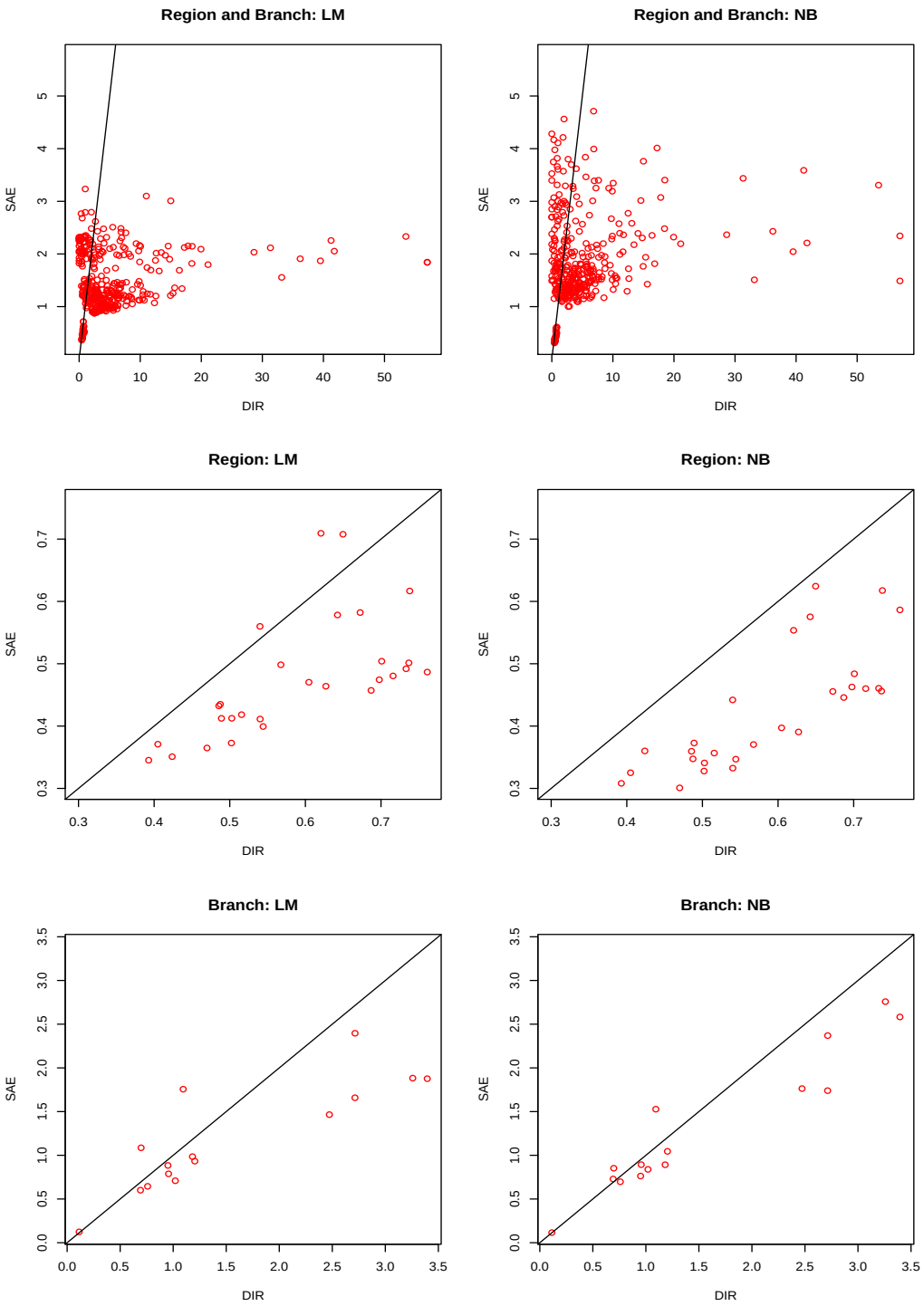

Figure F.3 Estimated standard error (SE) of the estimated mean number of absent days due to sickness at the region, the subbranch and their cross-classified domains levels based on the linear (LIN) and negative binomial (NB) multilevel models with the complex fixed effects component.

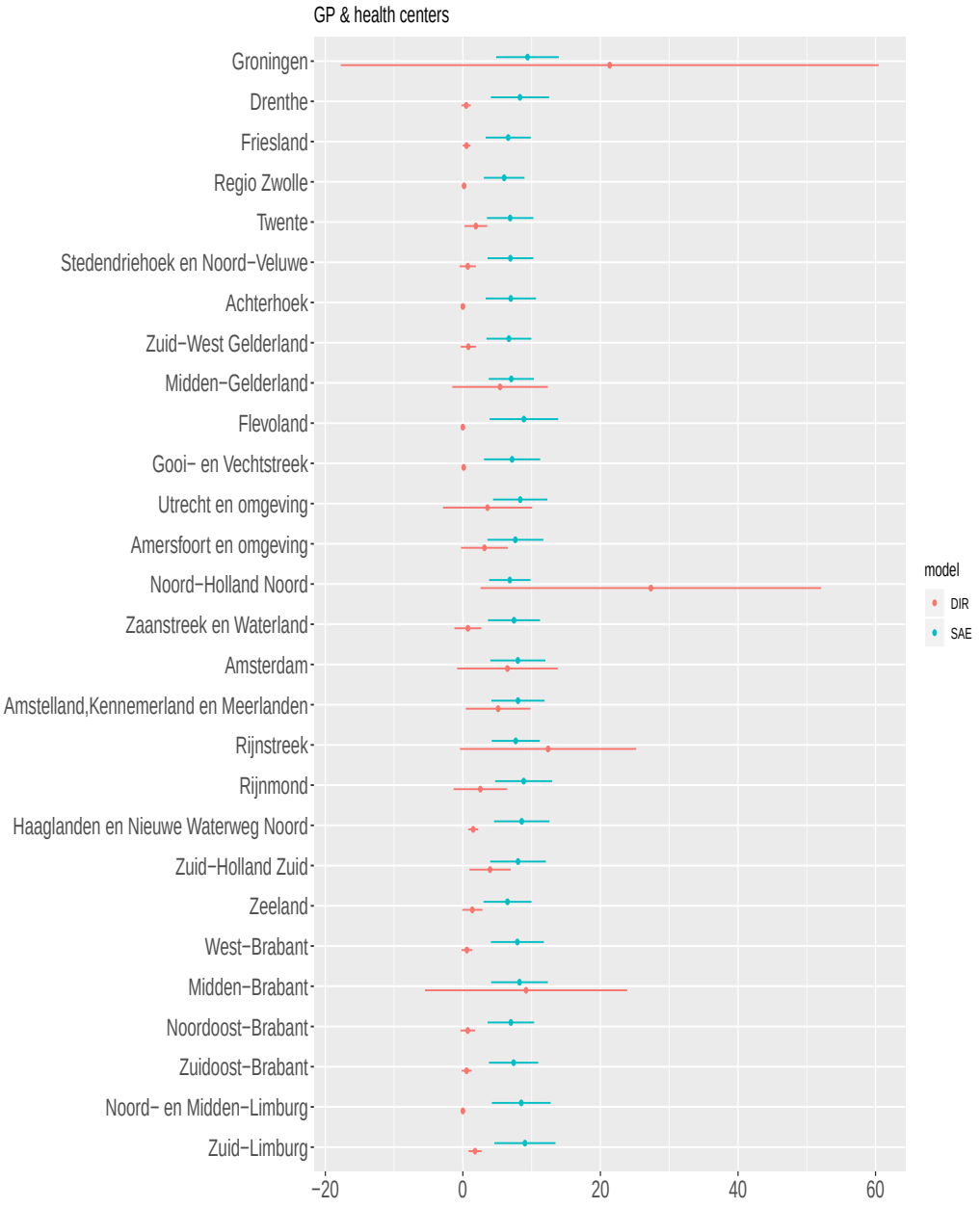

Figure F.4 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the number of absence days due to sickness at the region level for the "GP and health centers" subbranch.

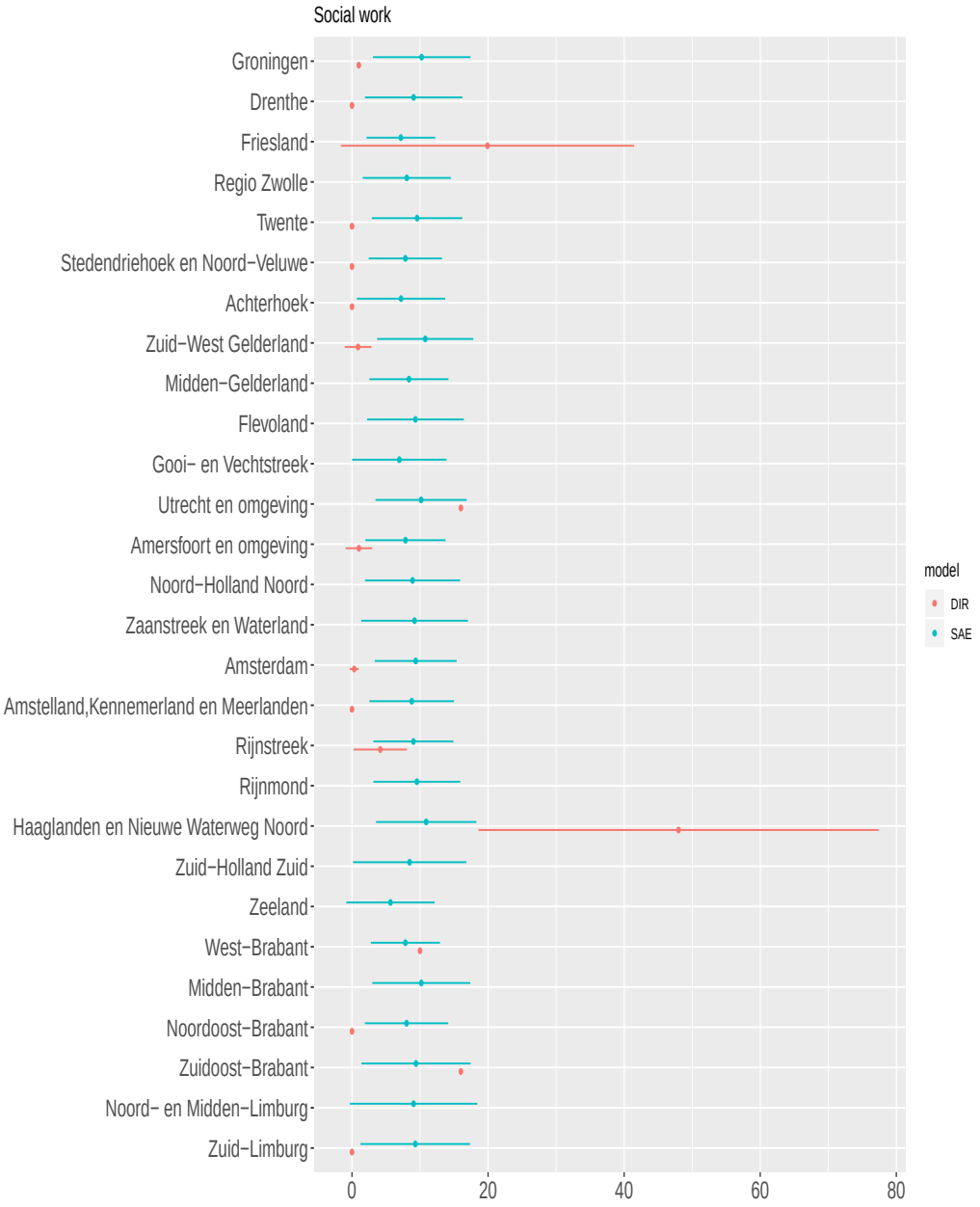

Figure F.5 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the number of absence days due to sickness at the region level for the "Social Work" subbranch.

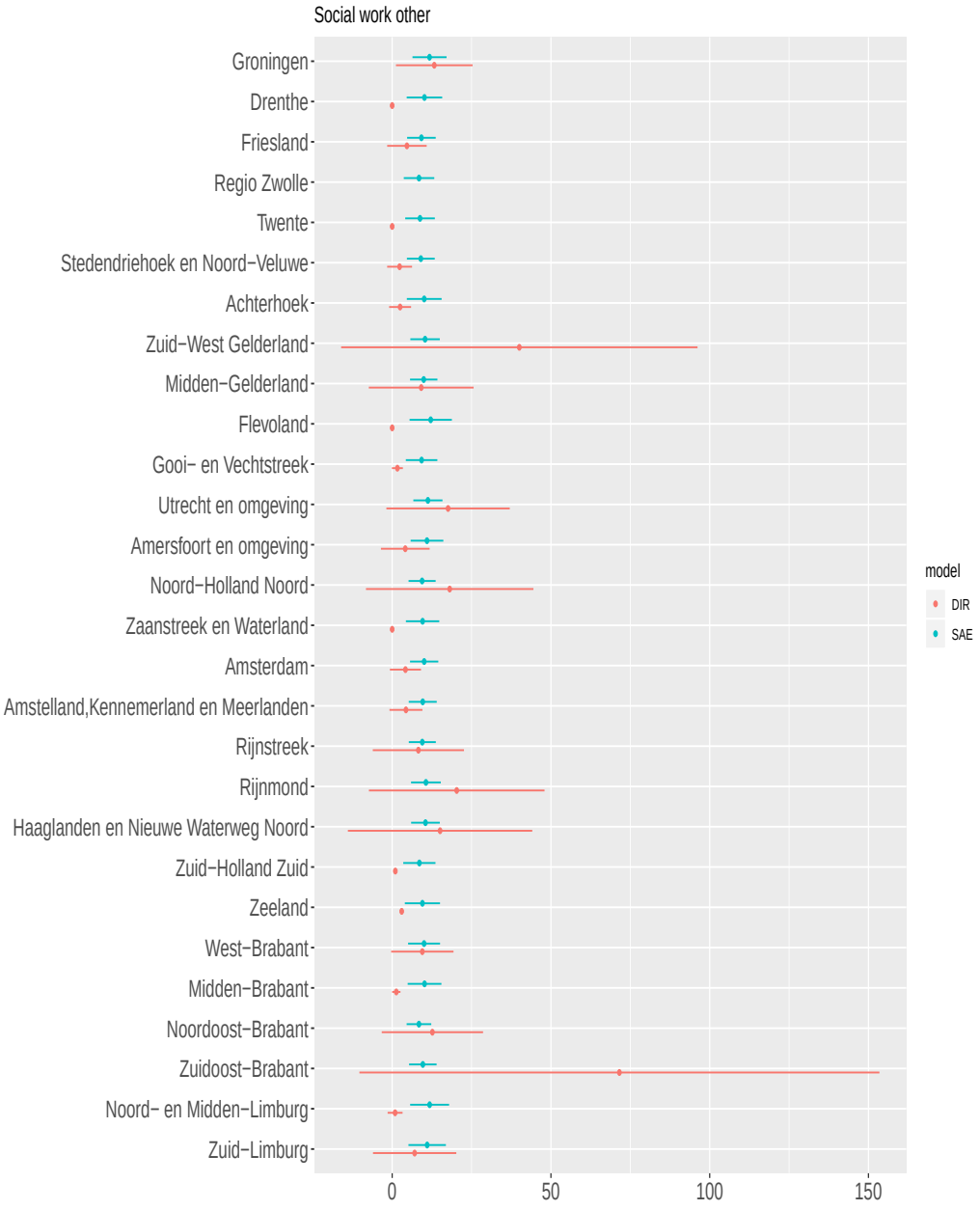

Figure F.6 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the number of absence days due to sickness at the region level for the "Other Social Work" subbranch.

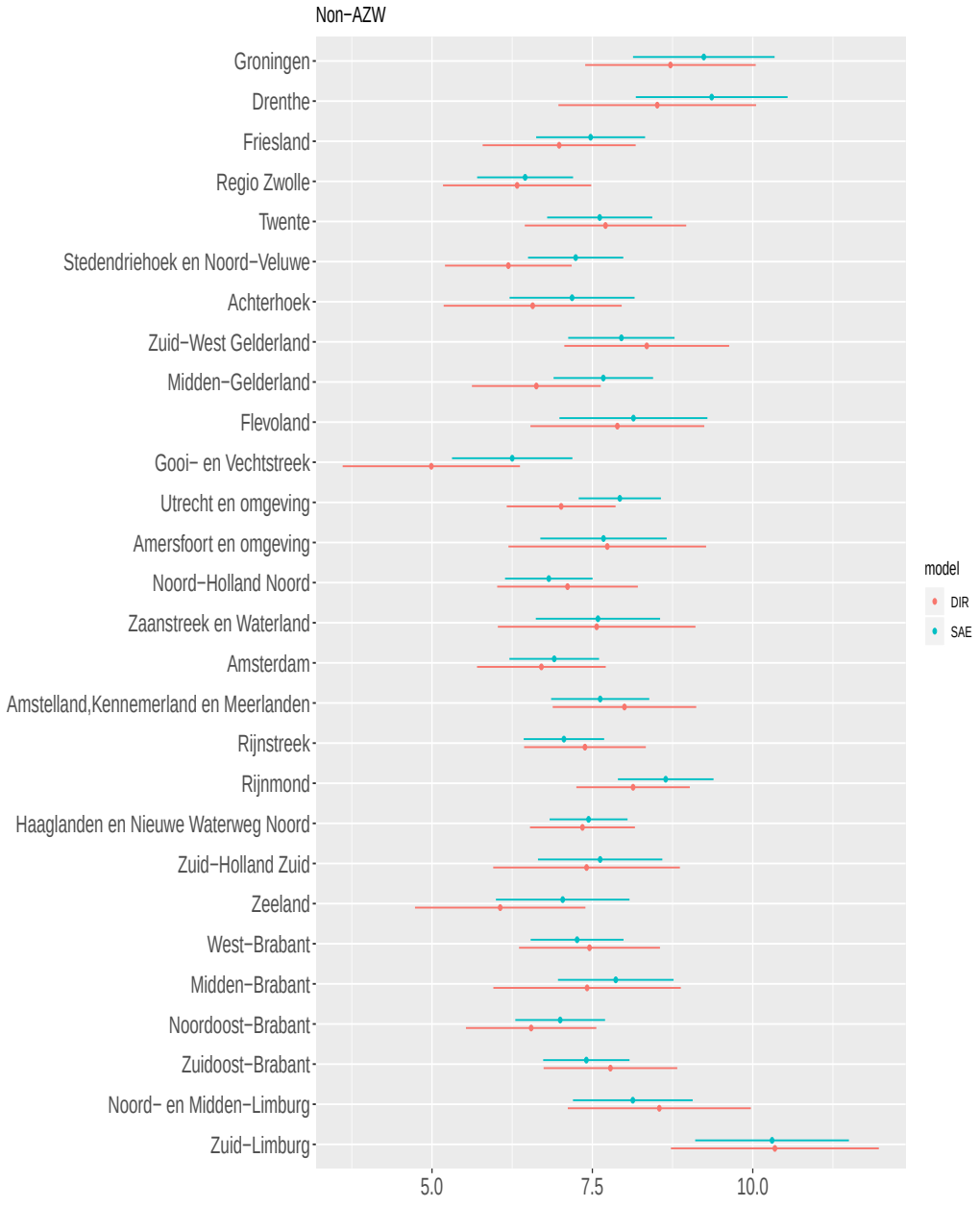

Figure F.7 Direct (DIR) and model-based (SAE) estimates with approximate 95% interval for the number of absence days due to sickness at the region level for the "Non-AZW" subbranch.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information