Adjusting for Miscounted Outcomes in Poisson and Negative Binomial Regression

Stefanie Muff, Milo A. Puhan and Leonhard Held

May 11, 2016

Measurement error in regression models has obtained a lot of attention in the past decades. However, error in count outcomes of Poisson or negative binomial regression has not been explicitly addressed in the literature. In this paper, we illustrate potential effects of miscounting error in the outcome by employing both theoretical considerations and simulations, and propose methodology to adjust for such error. While covariate-independent response error leads to potentially attenuated parameter estimates, covariate-dependent error may also overestimate the treatment effects. The overestimation of effects is, however, particularly delicate in the context of clinical trials, where an important assumption is that estimation procedures are conservative and do not overestimate treatment effects. A hierarchical Bayesian model, comprising negative binomial regression and error models, is suggested to account for covariate-dependent miscounting error. The methodology was applied to data from a large randomized controlled trial with the number of exacerbations of Chronic Obstructive Pulmonary Disease (COPD) patients as outcome. Validation data from an external study was used to obtain prior information on the response error model. Posterior distributions were estimated by a standard Markov chain Monte Carlo (MCMC) sampling approach. The results reveal that the treatment efficacy of the presented study might have been under- or overestimated due to miscounting error in the response.

1 Introduction

The modelling of measurement error (ME) and the description of its effects on the parameter estimates of regression models have a long tradition in the statistical literature (Pearson, 1902; Wald, 1940; Berkson, 1950; Fuller, 1987; Carroll et al., 2006; Buonaccorsi, 2010). The ME-induced biases in parameters can roughly be classified into attenuation (bias towards zero) and reverse attenuation (bias away from zero) effects. Interestingly, the vast majority of literature on ME in regression focusses on error in the covariates, which is also reflected by the attention given to it by recent monographs on error modelling (Gustafson, 2004; Carroll et al., 2006; Buonaccorsi, 2010).

In contrast to the covariates, which are not required to obey any distributional assumptions and are assumed to be error-free in standard regression methods, variability in the response is allowed and modelled via the likelihood of the regression model. The linear regression case, for instance, has been discussed by Abbrevaya and Hausman (2004) and Carroll et al. (2006), who mentioned that unbiased, additive, homoscedastic ME in the response of a linear model is simply absorbed in the variance of the distribution, and thus requires no additional modelling efforts. For heteroscedastic error in a continuous response, weighted regression or generalized least squares methods can be used (Carroll and Ruppert, 1988), while methods for the linear model with biased response were proposed by Buonaccorsi (1991, 1996) and Buonaccorsi and Tosteson (1993). On the other hand, there exists no variance term that could absorb response error in logistic regression, *i.e.* response misclassification, and this case thus needs specific treatment. Methods have been discussed in various papers (Ekholm and Palmgren, 1987; Copas, 1988; Neuhaus, 1999, 2002). Paulino et al. (2003) presented a Bayesian approach towards error modelling for the response in binomial regression, and Magder and Hughes (1997) proposed an EM algorithm to recover unbiased estimates of the odds ratios and their variances.

Our work was motivated by the paper by Frei et al. (2015), where it was recently shown that self-counted exacerbation numbers of COPD patients may contain considerable miscounting error. Exacerbation numbers are frequently used as a response variable in clinical trials. One example of such an application is the Towards a Revolution in COPD Health (TORCH) study (Calverley et al., 2007), where the rate of moderate COPD exacerbations was included as a secondary endpoint. Self-counted exacerbation numbers of COPD patients were used as proxies for the true values, and a negative binomial regression model was fitted. In this paper, we will therefore focus on *miscounting error* in count outcomes of Poisson or negative binomial regression models. Count outcomes occur in virtually all disciplines, for instance in ecology (White and Bennets, 1996), epidemiological studies (Wakefield, 2007) or in the context of randomized clinical trials (RCTs) (Suissa, 2006). Specific literature discussing miscounting error in the outcome of Poisson or negative binomial regression models is so far lacking. Effects of error in count outcomes of Poisson or negative binomial regression models

will be discussed here, and we propose a general modelling framework using a Bayesian approach to adjust for such error. We will formulate a hierarchical model, including a Poisson regression model and a negative binomial error model. In the simplest setup the error model is independent of the covariates, in which case the observed response is called a *surrogate* for the true response (Prentice, 1989; Carroll et al., 2006). However, it can be useful to formulate the response error models in more generality by allowing for covariate-dependency, as it was done for binomial regression (Neuhaus, 1999, 2002; Paulino et al., 2003). This leads to a *non-surrogate* response. The idea of a non-surrogate response is the analogue to the well-known concept of non-differential ME for error in a *covariate*, which holds if the error is independent of the response, given the true covariate(s) (Carroll et al., 2006). Implications of surrogate and non-surrogate responses and how to model them will be discussed, and potential effects will be illustrated from theoretical considerations and with simulations. In the application to the TORCH data, we use both a surrogate model and a model with an explicit dependency of the miscounting process on the binary treatment, and it will be shown that the results are sensitive to the model choice. A proper error model is thus crucial, as the correction might otherwise be in the wrong direction and then could lead to an increased bias. Our results indicate that the treatment-effect might have been underestimated in the original analysis.

An important step in error modelling, be it in covariates or in the response of regression models, is the estimation of the error model parameters from validation data, expert knowledge or instrumental variables (Carroll et al., 2006). Such information then enters in the form of prior information. However, hierarchical measurement error models are commonly nonidentifiable if all parameters are unknown, although Gustafson (2005) has illustrated that in a Bayesian setup, already the use of crude prior information may be an alternative to gaining identifiability through a change in the model formulation, in particular if there is sufficient indirect learning about nonidentifiable model parameters. However, we will show that crude priors are sometimes not sufficient in our models, which are very different from those considered by Gustafson (2005).

This paper is organized as follows. Section 2 recalls regression models for count outcomes and introduces the negative binomial error model for a miscounted response, which is then combined with the regression model to a hierarchical Bayesian model. Potential effects of error in the response are discussed. In Section 3 we provide some simulation results to illustrate the effect of miscounting response error in specific cases. We then apply the methodology in Section 4 to study data from COPD patients that were included in the TORCH study (Calverley et al., 2007). Finally, Section 5 provides some conclusions, and we discuss the importance, but also possible difficulties, of error modelling.

2 Methodology

2.1 Poisson and negative binomial regression models

Assume we have regression model with n observations, count outcome $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, a binary vector $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, hereafter called the *treatment*, and a covariate matrix $\boldsymbol{z}$ of dimension $n \times p$ containing p additional covariates. A count outcome y_i is often modelled with Poisson regression, $y_i \,|\, \mu_i \sim \text{Po}(\mu_i)$, where the mean $\mu_i = \mathsf{E}(y_i)$ and the linear predictor η_i are linked via

$$h_1^{-1}(\mu_i) \;\; = \;\; \eta_i \;=\; \beta_0 + \beta_x x_i + \boldsymbol{z}_i \boldsymbol{\beta}_z \;, \tag{1}$$

with link function h_1^{-1}, which is typically the log. The row vector $\boldsymbol{z}_i$ contains the p covariate measurements of individual i. The regression parameters β_0, β_x and $\boldsymbol{\beta}_z$ are the intercept, the treatment-effect and the parameters of the remaining covariates, respectively.

A possible limitation of the Poisson model is that the mean and the variance of the Poisson distribution are equal, which is an assumption that is often not justified in applications. Models with count outcome are thus frequently analyzed by negative binomial regression to account for additional variability between individuals, which results in overdispersion (Hilbe, 2011). A negative binomially distributed random variable $y \sim \text{NBin}(\mu, \theta)$ can be parameterized in various ways. Here, we use

$$\mathsf{Pr}(y = k) = \frac{\Gamma(\theta + k)}{k! \, \Gamma(\theta)} \left(\frac{\theta}{\theta + \mu} \right)^\theta \left(\frac{\mu}{\theta + \mu} \right)^k \;, \quad k = 0, 1, 2, \ldots \;,$$

with expected value $\mathsf{E}(y) = \mu$ and variance $\text{Var}(y) = \mu(1 + \mu/\theta)$. The parameter θ thus accounts for overdispersion, and Poisson data without overdispersion can be obtained for $\theta \to \infty$, *i. e.* the smaller θ, the more overdispersion. As in (1), the mean μ_i of a negative binomial regression

$$y_i \,|\, \mu_i \sim \text{NBin}(\mu_i, \theta) \tag{2}$$

is linked to the linear predictor via $\mu_i = h_1(\eta_i)$ with link function h_1^{-1} typically being the log or the identity link function.

2.2 Error model for count outcome

Assume that a count outcome y_i has been observed with error, so that not y_i, but a proxy $y_i^\star$ is recorded. Let us further assume that the observation $y_i^\star$ is again a count, *i. e.* that there is some miscounting error. It is then natural to formulate a negative binomial error model

$$y_i^\star \,|\, y_i \sim \text{NBin}(h_2(\gamma_0 + \gamma_1 y_i), \theta_E) \;, \tag{3}$$

with parameters $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^\top$ and overdispersion parameter $\theta_E > 0$. The function h_2 is the inverse link function, where h_2^{-1} is usually the identity function, but could be replaced with the log or any other monotonically increasing function. The identity link is predestined when the true and the observed counts are on the same scale.

At first glance it may be more natural to formulate the even simpler error model $y_i^\star \,|\, y_i \sim$ NBin(y_i, θ_E), $i.\,e.$ the special case with identity link and parameters $\gamma_0 = 0$ and $\gamma_1 = 1$. This is also the model that leads to unbiased error, $i.\,e.$ $\mathsf{E}(y_i^\star \,|\, y_i) = y_i$. However, outcomes with $y_i = 0$ then necessarily lead to observations $y_i^\star = 0$, meaning that zero counts are always correctly reported. This restriction seems unnatural, and it can be avoided only if $\gamma_0 > 0$ to allow for over-reporting in the case $y_i = 0$. This consideration shows that unbiasedness is not an essential property of count error modelling, as the error distribution is not naturally symmetrical, especially for small counts.

Note that the use of a negative binomial error model implicitly presumes that the error variance is $\mathrm{Var}(y_i^\star \,|\, y_i) \geq \mathsf{E}(y_i^\star \,|\, y_i)$, and equality holds when the model is Poisson, $i.\,e.$ $\theta_E = \infty$. The model thus imposes a minimal variance for the distribution of the observed counts around the true counts. In some situations such a modelling assumption could be implausible, in which case the negative binomial error model may be replaced by a count model that allows for underdispersion, such as the generalized event count model or the generalized Poisson distribution (Winkelmann, 2008). However, we do not further discuss underdispersion of count error models here.

The formulation of model (3) implies that $\mathsf{Pr}(y_i^\star \,|\, y_i, x_i, z_i) = \mathsf{Pr}(y_i^\star \,|\, y_i)$, $i.\,e.$ the error is independent of the covariates $(x_i, \boldsymbol{z}_i)$, given the true response y_i. Observation $y_i^\star$ is then called a *surrogate response* (Prentice, 1989; Carroll et al., 2006). In a more general setup, however, $y_i^\star$ may depend on the covariates $(x_i, \boldsymbol{z}_i)$. The most general formulation is then

$$y_i^\star \,|\, y_i, x_i, \boldsymbol{z}_i \sim \mathrm{NBin}(h_2(\gamma_0^{(x_i, \boldsymbol{z}_i)} + \gamma_1^{(x_i, \boldsymbol{z}_i)} y_i), \theta_E^{(x_i, \boldsymbol{z}_i)}) \ . \tag{4}$$

However, to keep notation simple, we omit the representation with covariate dependency from now on, except when explicitly needed.

2.3 The effect of a miscounted response

It is important to understand the effect of error-prone count outcomes $y_i^\star$ on the parameter estimates in the negative binomial regression model (2). As discussed in Section 2.2, the error model (3) may induce bias in the observed counts $y_i^\star$, $i.\,e.$ $\mathsf{E}(y_i^\star \,|\, y_i) \neq y_i$, and unbiased error is only obtained for the identity link $h_2^{-1} = \mathrm{id}$ and parameters $\gamma_0 = 0$ and $\gamma_1 = 1$. If the response error is unbiased, the parameter estimates for β_0 and β_x are also unbiased in the naive regression because $h_1^{-1}(\mathsf{E}(y_i^\star)) = h_1^{-1}(\mathsf{E}(y_i)) = \beta_0 + \beta_x x_i$. For $\gamma_0 = 0$ but $\gamma_1 \neq 1$, the error model is no longer unbiased, but with the standard log-link, the slope parameter

β_x can still be consistently estimated, as can be seen from

$$
\begin{aligned}
\log(\mathsf{E}(\mathsf{E}(y_i^\star \mid y_i))) &= \log(\mathsf{E}(\gamma_1 y_i)) \\
&= \log(\gamma_1) + \log(\mathsf{E}(y_i)) \\
&= \log(\gamma_1) + \beta_0 + \beta_x x_i \ .
\end{aligned} \tag{5}
$$

More generally, the likelihood for a response $y_i^\star$ following error model (4) can be written as

$$
\Pr(y_i^\star \mid x_i, \boldsymbol{z}_i) = \sum_{y_i} \Pr(y_i^\star \mid y_i, x_i, \boldsymbol{z}_i) \Pr(y_i \mid x_i, \boldsymbol{z}_i) \ . \tag{6}
$$

If $y_i^\star$ is a surrogate for y_i, the expression $\Pr(y_i^\star \mid y_i, x_i, \boldsymbol{z}_i)$ can be replaced by $\Pr(y_i^\star \mid y_i)$. If there is no relationship between y_i and the covariates $(x_i, \boldsymbol{z}_i)$, both terms of (6) are independent of $(x_i, \boldsymbol{z}_i)$, and thus also $y_i^\star$ is independent of the predictors. A naive regression analysis then leads to valid conclusions about the association of the predictors with the true response, although the resulting tests have decreased power, as discussed in Carroll et al. (2006)[section 15.4]. However, if the observation $y_i^\star$ is not a surrogate for y_i, there may be a relationship between $y_i^\star$ and $(x_i, \boldsymbol{z}_i)$, even if the true response does not depend on the covariates. Classical hypothesis tests for the regression parameters β_x and $\boldsymbol{\beta}_z$ are then no longer valid and their significance can be spurious (reverse attenuation). On the other hand, a true relation between the covariates and y_i may be masked by a surrogate *or* a non-surrogate response. The direction of a potential bias in the parameter estimates induced by the ME in $y_i^\star$ can thus not be predicted in general.

2.4 Hierarchical Bayesian model

Consider a negative binomial regression model with count outcome y_i and linear predictor η_i as given in equation (1), but assume that y_i can only be observed via a proxy $y_i^\star$ containing miscounting error. We then formulate a hierarchical model that comprises a Poisson model for the true observations, and a negative binomial error model:

$$
\begin{aligned}
y_i &\sim \ \mathrm{Po}\left(h_1(\eta_i)\right) \ , \\
y_i^\star \mid y_i &\sim \ \mathrm{NBin}\left(h_2(\gamma_0 + \gamma_1 y_i), \theta_E\right) \ .
\end{aligned} \tag{7}
$$

Again, h_1 and h_2 denote the inverse link functions of the regression and error models, respectively. The use of a Poisson regression model has an intuitive interpretation: all extra-variability in the measured response is attributed to the miscounting process. The assumption could be relaxed by using a negative binomial regression model, but it may then be difficult to identify the various contributors to the variance of $y_i^\star$, in particular if no prior information about the overdispersion parameter of the regression model is available. Independent normal priors with small precisions are usually specified for β_0, β_x and $\boldsymbol{\beta}_z$ in the

linear predictor η_i. Information on the parameters of the error model $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^\top$ and θ_E must be obtained from validation data or expert knowledge. If the error model is expected to be covariate-dependent, it is beneficial if the model parameters can be estimated from separate validation data (sub)sets.

The marginal distribution of the measured response $y_i^\star$ following model (7) is overdispersed by construction. This also holds if $y_i^\star \,|\, y_i$ is Poisson distributed, $i.\,e.$ when $\theta_E = \infty$, in which case the marginal expectation and variance of $y_i^\star$ are given as

$$
\begin{aligned}
\mathsf{E}(y_i^\star) &= \mathsf{E}(\mathsf{E}(y_i^\star \,|\, y_i)) \\
&= \mathsf{E}(h_2(\gamma_0 + \gamma_1 y_i)) \,, \\[1em]
\mathrm{Var}(y_i^\star) &= \mathsf{E}(\mathrm{Var}(y_i^\star \,|\, y_i)) + \mathrm{Var}(\mathsf{E}(y_i^\star \,|\, y_i)) \\
&= \mathsf{E}(h_2(\gamma_0 + \gamma_1 y_i)) + \mathrm{Var}(\mathsf{E}(y_i^\star \,|\, y_i)) \,.
\end{aligned}
$$

The last equality holds because variance and expected value are equal under the Poisson assumption. Except in the practically irrelevant case with $\gamma_1 = 0$, the term $\mathrm{Var}(\mathsf{E}(y_i^\star \,|\, y_i))$ is > 0, leading to $\mathrm{Var}(y_i^\star) > \mathsf{E}(y_i^\star)$, $i.\,e.$ overdispersion. Clearly, the variance term $\mathrm{Var}(y_i^\star)$ increases if the error model is overdispersed with $\theta_E < \infty$, in which case $\mathrm{Var}(\mathsf{E}(y_i^\star \,|\, y_i))$ (and thus the overdispersion) depends on all three error model parameters γ_0, γ_1 and θ_E. This indicates that strong prior information might be important in practical applications, as the parameters may otherwise not be identifiable.

The aspect of identifiability is a general concern in ME modelling, namely when the error model parameters are unknown (Gustafson, 2005). Equation (5), for instance, illustrates that confounding between γ_1 and β_0 could be an issue. Another situation that leads to potential identifiability problems is when the error model is treatment-dependent, $i.\,e.$ when there are two sets of error model parameters $\boldsymbol{\gamma}^{(0)} = (\gamma_0^{(0)}, \gamma_1^{(0)})^\top$ and $\boldsymbol{\gamma}^{(1)} = (\gamma_0^{(1)}, \gamma_1^{(1)})^\top$. The treatment effect β_x and the error model parameters are then confounded by construction, as both act on the difference between the treatment groups in the observed outcome $\boldsymbol{y}^\star$. Interestingly, Gustafson (2005) has illustrated that already relatively crude priors can be sufficient to obtain good results if there is enough indirect learning about nonidentifiable model parameters. However, it will be illustrated in Section 3 that crude priors may not be adequate for our models, in particular when the error model is treatment-dependent.

Irrespective of identifiability considerations, marginal posterior distributions for the parameters of model (7) can be obtained by Markov chain Monte Carlo (MCMC) sampling. Note that the latent variable $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is not Gaussian, and it is thus not possible to approximate the posterior marginals by integrated nested Laplace approximations (INLA), which are a computationally convenient alternative to sampling approaches for Bayesian inference in latent Gaussian models (Rue et al., 2009), in particular in the presence of covariate measurement error (Muff et al., 2015).

3 Simulation study

As discussed in Section 2.3, the effect of response ME as given in model (4) on the parameters in a negative binomial regression model cannot generally be predicted. In order to obtain some intuition of the biases induced by such error it is useful to generate data from different error models. Parameters estimates from naive regression models are then compared to their error-free counterparts, as well as to the results that are obtained when the error is modelled. The impact of nonidentified error models is also illustrated by incorporating both fixed error model parameters and prior uncertainty in the analyses of simulations 1 and 2. In each example presented here, we generated $n = 2000$ observations and included a treatment indicator $\boldsymbol{x}$ with $x_i = 0$ for $i = 1, \ldots, 1000$ and $x_i = 1$ otherwise. The linear predictor of regression model (2) thus simplifies to

$$\eta_i = \beta_0 + \beta_x x_i \ ,$$

with regression parameters fixed at simulation-specific values. Each simulation was repeated 200 times. In each iteration, parameter estimates for the treatment effect β_x from the following three models were stored:

 (i) the maximum likelihood (ML) estimate including the data without error.

 (ii) the ML estimate including the data with error, leading to the naive estimate.

(iii) the posterior mean of an MCMC sample for the hierarchical Bayesian model (7) with a burn-in of 1000 and a sampling of 5 000 iterations. Here, independent zero-mean Gaussian priors with a variance of 10^8 have been specified for β_0 and β_x.

ML estimates were obtained using the `glm.nb()` or `glm()` functions in `R` (R Core Team, 2013). MCMC samples were generated in JAGS via the `R`-interface `rjags` (Plummer, 2003). Means with 2.5% and 97.5% quantiles of the stored parameter estimates were then plotted in Figure 1.

Simulation 1: Surrogate response error
In the first example, data were generated as

$$
\begin{aligned}
y_i &\sim \mathrm{Po}\left(\exp(\beta_0 + \beta_x x_i), \theta\right) \ , \\
y_i^\star \,|\, y_i &\sim \mathrm{NBin}\left(\gamma_0 + \gamma_1 y_i, \theta_E\right) \ ,
\end{aligned}
\tag{8}
$$

with $(\beta_0, \beta_x) = (0.5, \log(0.65))$, $\boldsymbol{\gamma} = (0.4, 1.2)$, and $\theta_E = 4$. The effect of the erroneous response values $y_i^\star$ in naive regression was attenuation of β_x (bias towards 0), and thus the rate ratio $\exp(\beta_x)$ was biased towards 1, as illustrated in the upper left panel of Figure 1. Simple analytical calculations show that the error in the hierarchical model (8) always induces

this attenuation effect, independent of the error model parameters or the values of (β_0, β_x) (see Appendix B). In a first MCMC analysis, $\boldsymbol{\gamma}$ was included as fixed parameter vector using the same values as for generating the data, and the overdispersion parameter was given a log-normal prior $\theta_E \sim \mathrm{LN}(\log(4), 0.05)$. The posterior means were unbiased, but with larger variances than when the error-free data were used in an ML analysis. The results are labelled as Corrected.a in Figure 1.

A second analysis was then realized with the same prior for θ_E, but replacing the fixed parameters $\boldsymbol{\gamma}$ by a Gaussian prior centered around the correct values $\boldsymbol{\gamma} \sim \mathrm{N}((0.4, 1.2)^\top, \boldsymbol{\Sigma}(\boldsymbol{\gamma}))$ with covariance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \begin{pmatrix} 0.01 & -0.0025 \\ -0.0025 & 0.01 \end{pmatrix} . \tag{9}$$

This reflects a prior standard deviation of 0.1 for γ_0 and γ_1 with prior correlation of -0.25. The results labelled as Corrected.b indicate that $\exp(\beta_x)$ could still be estimated with almost no bias, despite the use of flexible priors. To further study the effect of prior uncertainty of the $\boldsymbol{\gamma}$ parameters on the estimate of the rate ratio, the analysis was repeated with the same setup, but replacing the covariance matrix (9) with $10 \cdot \boldsymbol{\Sigma}(\boldsymbol{\gamma})$. The results labelled as Corrected.c then show that the estimate moves further away from the correct value, although the bias is not too severe. However, this example illustrates that identifiability considerations are relevant in the hierarchical models presented here, as already discussed in Section 2.4. In contrast to the models approached in Gustafson (2005), the lesson from this simulation is that prior uncertainty in the error model parameters should be incorporated with great care, and vague priors should be avoided, if possible.

Simulation 2: Non-surrogate error with treatment-dependent modelling
We again used the log link for the Poisson regression model, and $(\beta_0, \beta_x) = (0, \log(0.65))$. A treatment-dependent negative binomial error model

$$y_i^\star \mid \{y_i, x_i = j\} \sim \mathrm{NBin}\left(\gamma_0^{(j)} + \gamma_1^{(j)} y_i, \theta_E^{(j)}\right) , \quad j = 0, 1 , \tag{10}$$

was used with $\boldsymbol{\gamma}^{(0)} = (0.4, 0.8)^\top$, $\boldsymbol{\gamma}^{(1)} = (0.2, 1.2)$, and $\theta_E^{(0)} = \theta_E^{(1)} = 4$. The naive estimate of $\exp(\beta_x)$ was biased towards 1, *i.e.* attenuated, as can be seen in the upper right panel of Figure 1.

The treatment-dependency was incorporated by assuming $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(1)}$ to be known. The dispersion parameters $\theta_E^{(0)}$ and $\theta_E^{(1)}$ were given log-normal priors $\mathrm{LN}(\log(4), 0.05)$. The results from the MCMC simulations are depicted as Corrected.a in Figure 1, showing that correct parameter estimates can be recovered. We then added separate Gaussian priors for $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(1)}$, centered around the correct values and with covariance matrix (9). The results labelled as Corrected.b indicate that almost unbiased parameter estimates for $\exp(\beta_x)$ could be recovered, despite the incorporation of some uncertainty in the priors.

The analysis was again repeated with ten times larger variances and covariances on the $\boldsymbol{\gamma}$ priors. The results labelled with Corrected.c show that the model is then unable to recover approximately correct parameter estimates on average. The treatment effect is on average more attenuated than in the naive approach, and the variability has increased dramatically. This result underlines the identifiability problem of the hierarchical model with treatment-dependent error, as discussed in Section 2.4. We conclude that prior uncertainty in treatment-dependent error model parameters should be incorporated with even more care than for surrogate error models. The use of informative priors should be considered, if possible.

Simulation 3: Non-surrogate error with surrogate modelling

The aim of this simulation was to illustrate how a treatment-dependent response error may lead to reverse attenuation and thus to overestimated treatment-effects. The error model was given as in equation (10) with $(\beta_0, \beta_x) = (0.5, 0)$, $\boldsymbol{\gamma}^{(0)} = (0, 1.2)^\top$, $\boldsymbol{\gamma}^{(1)} = (0.2, 0.8)^\top$, and $\theta_E^{(0)} = \theta_E^{(1)} = 4$. Note that x_i is now chosen to have no effect on the response, since $\beta_x = 0$. The naive estimate of $\exp(\beta_x)$ was severely affected by the error in the response, as can be seen in the bottom left panel of Figure 1.

In contrast to simulation 2, the naive effect appeared stronger, *i. e.* the estimated rate ratio was biased away from 1, as can be seen in the plot labelled as Naive in Figure 1. This shows that the bias induced by treatment-dependent error in the response may be in both directions. In practice this would mean that the naive analysis does not give a conservative estimate of the treatment effect. To illustrate that an error model that ignores a treatment-dependency is generally not able to recover unbiased parameter estimates, we incorporated a surrogate response model for $y_i^\star \,|\, y_i$ using error parameters fixed at the true value $\theta_E = 4$ and at the average values of the two groups $\boldsymbol{\gamma} = (0.1, 1.0)^\top$. The correct value $\exp(\beta_x) = 1$ could not be recovered, and although the correction was in the right direction, the use of such average error model parameters appears essentially useless. The results for this case are labelled as Corrected in the bottom left panel of Figure 1.

Simulation 4: Surrogate error with wrong parameters

Suitable prior information on the error model parameters is important to obtain useful corrections of regression parameters. To illustrates this aspect of response error modelling, we generated data according to model (8) with exactly the same parameters as in simulation 1 with $(\beta_0, \beta_x) = (0.5, \log(0.65))$, $\boldsymbol{\gamma} = (0.4, 1.2)^\top$ and $\theta_E = 4$. Error correction was then done by using a misspecified error model with fixed values $\boldsymbol{\gamma} = (0, 1.2)^\top$, *i. e.* with a wrong γ_0 parameter. The overdispersion parameter was again given the log-normal prior $\theta_E \sim \mathrm{LN}(\log(4), 0.05)$. Instead of diminishing the bias, the resulting correction then led to estimated rate ratios that were even closer to 1, meaning that the bias was enlarged. The results are depicted in the bottom right panel of Figure 1.
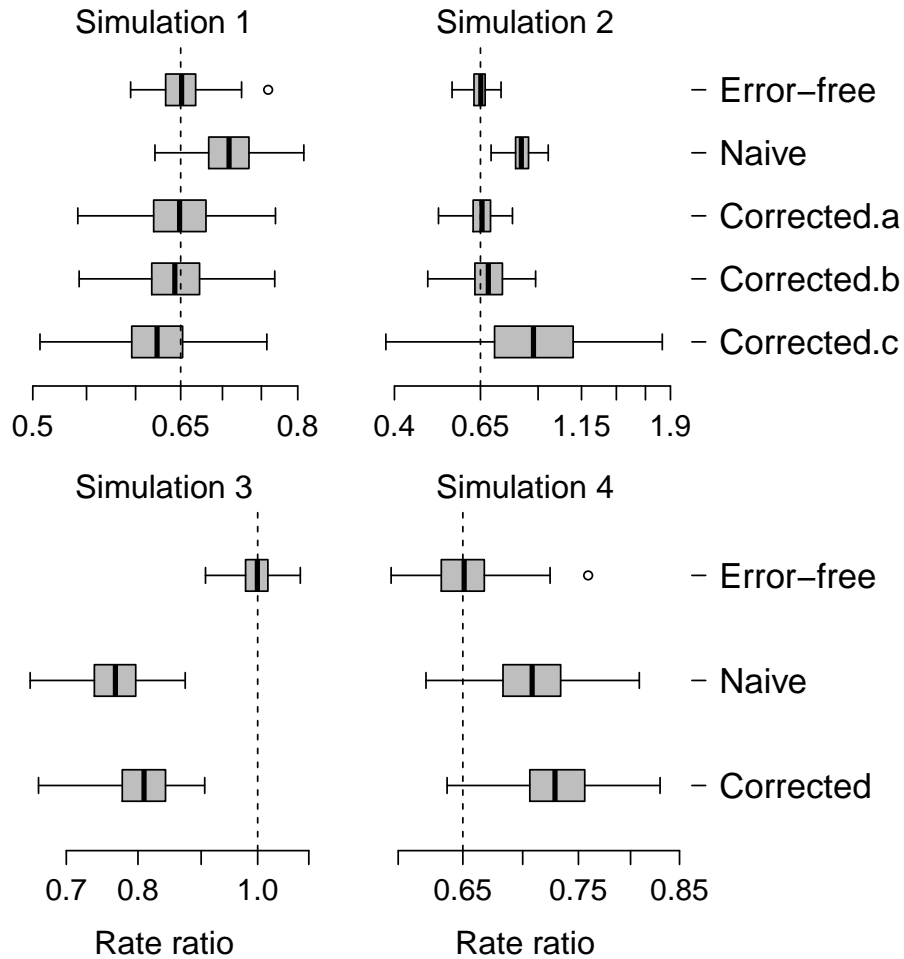
Figure 1: Boxplots for the ML estimates of error-free and naive estimates, as well as for the posterior means for the error-corrected estimates of the treatment vs. placebo rate ratio $\exp(\beta_x)$. Each boxplot was generated from the 200 iterations of the simulations. Note that the x-axis is given in log-scale.

4 Application to the TORCH study data

The TORCH trial (Calverley et al., 2007) was a large trial of pharmacotherapy in patients with COPD that lasted for 3 years. The study included 6112 patients in the efficacy population, of which $n_1 = 1524$ received a placebo and $n_2 = 1533$ a combination treatment (salmeterol plus fluticasone). Another 1521 patients received only salmeterol and 1534 received only fluticasone, but these treatment arms were not included in our analyses. The primary objective of the study was to demonstrate a significant reduction in all-cause mortality in COPD subjects that obtained the combination treatment, compared with the placebo group. The rate of moderate COPD exacerbations was included as a secondary endpoint, which is the interest of this example. The data from the TORCH study was provided by a data access system and was accessed through the SAS Solutions on Demand secure portal (`https://researchenvadmin.ondemand.sas.com`). All analyses were carried out in the provided Clinical Trial Data Transparency Research Environment, from where results could then be exported.

Calverley et al. (2007) analyzed the frequency of exacerbations using a negative binomial model, with $\boldsymbol{x}$ indicating the treatment ($x_i \in \{0, 1\}$), adjusted for region of recruitment (Eastern Europe, Western Europe, USA, Asia and Pacific, Other), age, sex, baseline smoking status (yes/no), BMI, number of exacerbations in the 12 months prior to screening (categorized as 0, 1, ≥ 2), and baseline disease severity. These confounder variables were summarized in the matrix $\boldsymbol{z}$. To account for inter-individual differences in the time under treatment, the logarithm of the time $\log(E_i)$ during which individual i received treatment was included as an offset variable (Suissa, 2006). The rate ratio for treatment vs. placebo was estimated as 0.75 with a 95% confidence interval of $(0.69, 0.81)$. These values are depicted as the "naive" result in Figure 2. Note that the presence of additional covariates introduces another level of complexity, compared to the simulations of Section 3 without covariates apart from $\boldsymbol{x}$.

Importantly, the exacerbation numbers that were used in the analysis of Calverley et al. (2007) stem from patient self-reports, and miscounting error is thus a concern. Denote by $\boldsymbol{y}^\star = (y_1^\star, \ldots, y_n^\star)$ the vector of self-reported exacerbation numbers, with $y_i^\star$ being the number reported by patient i. We modelled the error in $\boldsymbol{y}^\star$ using the data from Frei et al. (2015), an external validation study, where self-reported exacerbation numbers of COPD patients were compared to the numbers ascertained by an adjudication committee who had access to the patients charts of their general practitioners, patient self-reports and from all follow-up assessments. The aggregated validation data are shown in Table 1. A negative binomial regression model with identity link was fitted to describe the distribution of the reported counts as a function of the true counts. The model parameters were estimated without stratification for treatment or any additional covariates, which seemed natural, as the validation data in Frei et al. (2015) stem from an observational study and were thus not

collected in the context of a randomized clinical trial.

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	127	24	5	4	2	2	1	0	0	0	0	0	0
1	26	40	5	2	1	3	0	0	0	0	0	0	0
2	9	17	10	4	2	1	0	0	0	0	0	0	0
3	3	6	7	10	2	3	2	1	0	0	0	0	0
4	1	7	3	6	2	3	2	1	0	0	0	1	0
5	0	3	5	4	0	4	1	1	0	0	0	0	0
6	0	2	4	1	6	1	2	0	0	0	0	0	0
7	0	2	2	0	2	0	0	0	0	0	0	0	0
8	0	0	0	2	2	0	1	2	1	0	0	0	1
9	0	0	0	1	0	0	0	1	1	0	0	0	0
10	0	0	0	0	0	1	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	2	0	0	2	0	0	0	0
13	0	0	0	0	0	1	1	0	0	0	0	0	0
14	0	0	0	0	1	0	0	0	0	0	1	0	0
15	0	0	0	0	0	0	0	0	0	0	0	1	0

Table 1: Data on miscounting. Shown is the total number of centrally adjudicated exacerbations per patient (columns) by the total number of self-reported exacerbations per patient (rows).

ML estimates and variance-covariance matrix for the error model parameters $\boldsymbol{\gamma}$ are given as

$$
\begin{aligned}
\hat{\boldsymbol{\gamma}} &= (0.3644, 1.0762)^{\top} , \\
\boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}}) &= \begin{pmatrix} 0.00229 & -0.000989 \\ -0.000989 & 0.00460 \end{pmatrix} ,
\end{aligned}
\tag{11}
$$

and the overdispersion parameter was estimated as $\hat{\theta}_E = 3.49$ with a standard error of 0.765. The fact that $\hat{\theta}_E \ll \infty$ indicates that an error model with overdispersion is appropriate for the miscounting error in this study. It was checked that the residuals are in the expected range without obvious outliers. The plot with the deviance residuals versus fitted values is shown in Figure 3 of Appendix A.

We started by formulating a hierarchical model for the analysis of the TORCH trial, including a surrogate response error model as

$$
\begin{aligned}
y_i &\sim \mathrm{Po}\left(\exp(\log(E_i) + \beta_0 + \beta_x x_i + \boldsymbol{\beta}_z^{\top} \boldsymbol{z}_i)\right) , \\
y_i^{\star} \,|\, y_i &\sim \mathrm{NBin}\left(\gamma_0 + \gamma_1 y_i, \theta_E\right) .
\end{aligned}
\tag{12}
$$

Independent $\mathrm{N}(0, 10^2)$ priors were assigned to the components of $\boldsymbol{\beta} = (\beta_0, \beta_x, \boldsymbol{\beta}_z^{\top})^{\top}$. The overdispersion parameter of the error model, estimated as $\hat{\theta}_E = 3.49$, was given a log-

normal prior $\mathrm{LN}(\log(3.49), 0.22^2)$, whereas the second argument is the squared standard error that was calculated using the delta-rule for log-transformed variables, $\mathrm{se}(\log(\hat{\theta}_E)) = 0.765/3.49 = 0.22$. A Gaussian distribution with moments estimated in (11) was taken as prior for $\boldsymbol{\gamma}$, because simulation 1 indicated that the treatment-effect β_x is not severely affected by confounding problems when flexible priors are used. Importantly, by employing the parameter estimates from the validation data as prior information for the TORCH analysis, we implicitly assume that the error model deduced from the validation data is *transportable* to the TORCH study. This seems to be a plausible assumption here, as both data sets were collected over the same duration of three years, and the mean number of reported exacerbations was 2.02 in the validation data, while it was 2.19 in the TORCH study, thus length and exacerbation counts were similar.

Independent normal priors $\mathrm{N}(0, 10^2)$ were assigned to the components of $\boldsymbol{\beta} = (\beta_0, \beta_x, \boldsymbol{\beta}_z^\top)^\top$. Following the discussion in Section 2.4, we used a Poisson model for y_i by setting $\theta = \infty$. The overdispersion parameter of the error model, that has been estimated as $\hat{\theta}_E = 3.49$, was given a flexible $\mathrm{U}[0,100]$ prior.

I'm not happy with the uniform prior on θ_E. Better using a gamma prior? Anything else?

A Bayesian analysis using MCMC was performed in `rjags` by running two parallel chains for $25\,000$ iterations each, with a burn-in of $2\,500$ and a saving frequency of 5. The posterior mean of the rate ratio for treatment vs. placebo was $\exp(\hat{\beta}_x) = 0.81$ with a 95% credible interval (CI) from 0.73 to 0.89. The graph labelled as Corrected.a in Figure 2 depicts this estimate in comparison to the uncorrected estimate from Calverley et al. (2007). Error-correction hence leads to a less pronounced estimated treatment effect, but also to more uncertainty. To ascertain that the estimate of the treatment effect was not biased by a confounding effect originating from the flexible priors on the error model parameters (see simulations 1 and 2), we repeated the calculation by using fixed values $\boldsymbol{\gamma}^\top = (0.3644, 1.0762)$. This did not change the result for $\exp(\hat{\beta}_x)$, which was then 0.81 with a slightly narrower 95%-CI ranging from 0.74 to 0.89.

It has been illustrated in simulation 1 and analytically derived in Appendix B that surrogate response error always leads to attenuation effects, thus reverse attenuation can only occur when the error is treatment-dependent. The fact that surrogate error modelling led to a correction towards 1 thus indicates that the model did not correctly capture the error structure, or that the error parameters were not correctly specified, see simulation 4. Because information on the treatments of the patients in the validation data was available, we incorporated a treatment-dependency into the error model. Simulation 2 revealed that, due to identifiability problems, treatment-dependent error modelling is particularly challenging and requires the specification of informative priors for $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(1)}$. To obtain plausible prior information on $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(1)}$, we split our validation data set into two groups: one group with patients that received a pharmacotherapy comparable to the one investigated in the TORCH study ($n_1 = 332$ patients), and a second group that did not ($n_0 = 75$ patients).

Estimates

$$
\hat{\boldsymbol{\gamma}}^{(0)} = (0.3973, 0.8433)^\top, \quad \boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}}^{(0)}) = \begin{pmatrix} 0.00871 & -0.00464 \\ -0.00464 & 0.03024 \end{pmatrix}, \tag{13}
$$

$$
\hat{\boldsymbol{\gamma}}^{(1)} = (0.3550, 1.1015)^\top, \quad \boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}}^{(1)}) = \begin{pmatrix} 0.00309 & -0.00130 \\ -0.00130 & 0.00536 \end{pmatrix}, \tag{14}
$$

were obtained when fitting the negative binomial regression models with identity link to the two subsets. The overdispersion parameters were estimated as $\hat{\theta}_E^{(0)} = 6.36\,(\mathrm{se} = 7.19)$ and $\hat{\theta}_E^{(1)} = 3.39\,(\mathrm{se} = 0.77)$. We first modelled the error using values fixed at the point estimates from (13) and (14) for $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(1)}$, log-normal priors $\theta_E^{(0)} \sim \mathrm{LN}(\log(6.36), 1.13^2)$ and $\theta_E^{(1)} \sim \mathrm{LN}(\log(3.39), 0.23^2)$, while keeping the priors for $\boldsymbol{\beta}$ as before. The posterior mean treatment effect from the MCMC simulations was then $\exp(\hat{\beta}_x) = 0.62$ with 95%-CI ranging from 0.56 to 0.69. This result is depicted as Corrected.b in Figure 2. The use of a treatment-dependent response error model thus led to a noticeable increase in the treatment effect estimate, although its uncertainty is probably underestimated due to the use of fixed error model parameters $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(1)}$.

A final analysis was then conducted by including flexible priors on the error parameters, using the covariance matrices from (13) and (14). Note, however, that the prior covariance matrix for $\boldsymbol{\gamma}^{(0)}$ in (13) suffers from more uncertainty than the one for $\boldsymbol{\gamma}^{(1)}$ in (14). This large uncertainty is at least partially due to the low number of patients in the group without treatment ($n_0 = 75$ vs. $n_1 = 332$). On the other hand, it must be kept in mind that the validation data were not collected in the context of an RCT, meaning that patients without treatment are probably those with less severe symptoms and thus experienced fewer exacerbations, which also leads to more uncertainty in the parameter estimates. The estimate $\exp(\hat{\beta}_x) = 0.74$ and 95%-CI ranging from 0.53 to 1.03 are depicted as Corrected.c in Figure 2. The shift of the estimate towards 1 and the increase of the CI when changing from fixed to flexible priors for the $\boldsymbol{\gamma}$ parameters is exactly what has been observed in simulation 2, $i.\,e.$ a bias due to nonidentifability of the model. We thus conclude that the results for fixed error model parameters are more trustworthy.

It would be tempting to use the Gaussian priors according to error model (11) for $\boldsymbol{\gamma}$. However, by doing this we would implicitly assume that the error model deduced from the validation data is transportable to the TORCH study. Such an assumption can, however, lead to a prior-data conflict (Box, 1980), particularly if the conditions under which the validation data were collected differ from the study conditions, which is difficult to verify. We therefore multiplied the covariance matrix from the validation data with a scalar $g > 0$, leading to

$$
\boldsymbol{\gamma} \sim \mathrm{N}(\hat{\boldsymbol{\gamma}}, g\boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}})), \tag{15}
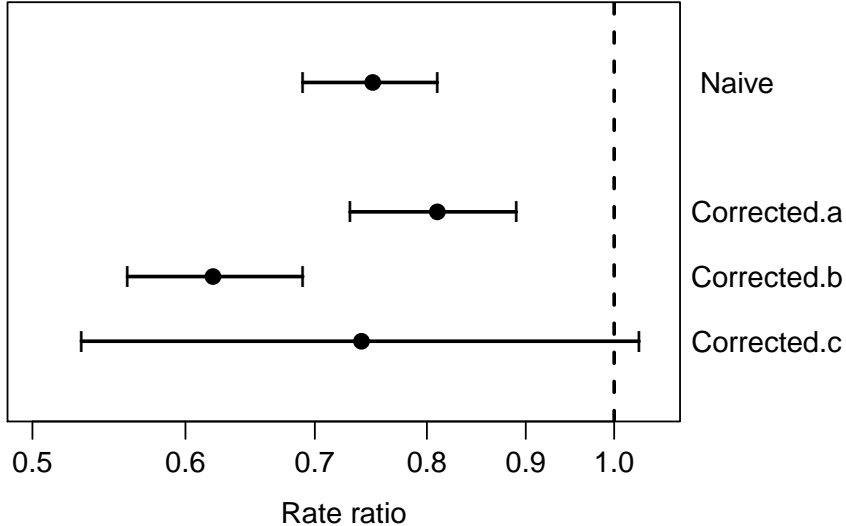$$

Figure 2: Naive and error-corrected estimates for the treatment vs. placebo rate ratio $\exp(\beta_x)$ of exacerbation rates in the TORCH trial. Horizontal lines represent the 95% confidence/credible intervals. The middle graph contains the result for the case when a surrogate error model was used (Corrected.a), while the third and fourth graphs shows the results for the treatment-dependent models with fixed (Corrected.b) and flexible (Corrected.c) priors on $\gamma^{(0)}$ and $\gamma^{(1)}$. The vertical dashed line indicates a rate ratio of 1, *i.e.* no treatment effect. Note that the x-axis is given in log-scale.

and used a hyper-g prior (Liang et al., 2008; Held et al., 2015) with

$$t = \frac{g}{g+1} \sim \mathrm{U}(0,1) \; . \tag{16}$$

A Bayesian analysis using MCMC was performed in `rjags` by running two parallel chains for $25\,000$ iterations each, with a burn-in of $2\,500$ and a saving frequency of 5. The posterior mean of the rate ratio for treatment vs. placebo was $\exp(\hat{\beta}_x) = 0.81$ with a 95% credible interval of $(0.73, 0.89)$, which corresponds to a posterior probability $\mathsf{Pr}(\exp(\beta_x) \geq 1 \,|\, \mathrm{data}) < 0.0001$. The middle graph in Figure 2 depicts this estimate in comparison to the uncorrected estimate from Calverley et al. (2007) ($\exp(\hat{\beta}_x) = 0.75$, 95%-CI $(0.69, 0.81)$). Error-correction hence leads to a less pronounced estimated treatment effect, but also to more uncertainty. Importantly, the posterior median of g was 5.04 with equi-tailed 95% credible interval ranging from 0.87 to 38.44. The hyper-g prior thus decreased the weight of the prior on the regression coefficients for the error model by a median factor of $1/5.04 = 0.20$ (95% CI: 0.03 to 1.15), which indicates that the error model deduced from Frei et al. (2015) indeed is in some conflict with the error structure in the count response of the TORCH study, although the uncertainy in g is large. Finally, the posterior median of θ_E is 3.59 with 95%-CI 2.68 to 5.31, which is very close to the value observed in the validation study.

It has been illustrated in Section 3 that reverse attenuation effects may be the result

of a treatment-dependent response error model. However, as already mentioned, there is no such information available here, neither from the study data (where miscounting error was not assessed), nor from the validation data (where no treatments were involved). It is nevertheless possible to introduce a treatment-dependency into model (12) by replacing the error model with its treatment-dependent version, identical to the one formulated in (10). This model was implemented here by using the same priors as for the surrogate response error model for both sub-models, i.e., $\boldsymbol{\gamma}^{(0)} = (\gamma_0^{(0)}, \gamma_1^{(0)})^\top$ and $\boldsymbol{\gamma}^{(1)} = (\gamma_0^{(1)}, \gamma_1^{(1)})^\top$ were given identical and independent priors

$$\begin{pmatrix} \boldsymbol{\gamma}^{(0)} \\ \boldsymbol{\gamma}^{(1)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \hat{\boldsymbol{\gamma}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix}, \begin{pmatrix} g_0 \boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}}) & \mathbf{0} \\ \mathbf{0} & g_1 \boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}}) \end{pmatrix} \right) ,$$

where $\mathbf{0}$ is the 2×2 matrix of zeros, assuming that the error model parameters for the treatment and the placebo groups are independent. Individual inverse prior weights g_0 and g_1 for the two treatment-groups were used, and both were given the same hyperpriors as in (16). The priors for $\boldsymbol{\beta}$, θ, $\theta_E^{(0)}$ and $\theta_E^{(1)}$ were left unchanged. The posterior mean treatment effect was now $\exp(\hat{\beta}_x) = 0.97$ with a credible interval of $(0.63, 1.64)$. The posterior probability $\mathsf{Pr}(\exp(\beta_x|\text{data}) \geq 1) = ...$, thus indicating even less evidence for the efficacy of the therapy. The third level of Figure 2 illustrates this graphically. Again of interest are the posterior median values for $w_0 = 1/g_0$ and $w_1 = 1/g_1$. While w_0 was estimated to be 0.081 (95%-CI: 0.007 to 0.513) and thus considerably downweighted the prior for the placebo group, the posterior median of w_1 was 7.22 (95%-CI: 0.010 to 45.895), and thus increased the weight of the prior distribution, although the uncertainty in w_1 was large.

check from simulation results

Figure **??** shows as lines the priors for the error model parameters, in comparison to their posterior marginals, which are represented by histograms. The Figure clearly shows two things: first, the prior and posterior distribution for $\boldsymbol{\gamma}^0$ are in strong disagreement, indicating a prior-data conflict, which is also reflected by the small posterior median of w_0. On the other hand, priors and posteriors for $\boldsymbol{\gamma}_1$ are in good agreement, which is reflected by the large posterior median of w_1. Second, the substantial differences in the posteriors of $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(1)}$ further strengthens the conjecture that treatment-dependent error modelling is necessary here.

5 Discussion

The use of a miscounted response in a Poisson or negative binomial regression may lead to attenuated or reversely attenuated parameter estimates. We have proposed a statistical framework to treat error in count outcomes to recover unbiased estimates of the regression parameters. To this end, a negative binomial error model was formulated, and we have shown how a Bayesian approach can be employed to jointly estimate posterior marginals for the hierarchical model that encompasses a regression and an error model.

In its most general form, the observed counts depend on the true counts via a covariate-dependent error model, and it has been discussed in Section 2.3 and illustrated in Section 3 that the direction of the bias in the parameters is then not generally predictable. On the other hand, if the count error is independent of the regression covariates, $i.\,e.$ if $\boldsymbol{y}^\star$ is a surrogate response for $\boldsymbol{y}$, slope parameters are either unbiased or attenuated. In this case, classical statistical tests that evaluate the predictive ability of a covariate are still valid, albeit less powerful (Carroll et al., 2006). Most applied researchers are aware of biases induced by ME in covariates or in the response of regression models. However, the assumption that the observed covariates or the response are a surrogate for the true values is often made, in which case the estimated effects are considered to be a lower bound of the true effects.

The proposed method has been applied here to data from the TORCH study, where the efficacy of a treatment on the exacerbation rate of COPD patients was studied (Calverley et al., 2007). Exacerbation numbers from patient self-reports have been included as response values in the negative binomial model of this study, but these values have recently been shown to contain considerable error (Frei et al., 2015). A Bayesian hierarchical model, including a Poisson regression model and a negative binomial error model, was formulated, and posterior inference was done using MCMC. We started by employing a treatment-independent response error model, which led to a smaller estimated treatment effect than in the original TORCH study. The fact that surrogate error modelling corrected the treatment effect $\exp(\beta_x)$ towards 1 indicated that the error model was misspecified, because a correctly specified error model of a covariate-independent error structure would lead to a correction in the opposite direction. It is, however, not clear whether the condition of covariate-independence is violated, or wheter the parameters of the surrogate model were misspecified, as it was the case in simulation 4. To illustrate that a change of the modelling assumptions can have a strong effect on the results, we allowed for treatment-dependency in the error model and used prior information from our validation data set, which resulted in a considerably more pronounced treatment effect estimate. The TORCH example thus illustrated two important points: first, the choice of the error model is crucial. We could argue why the surrogate model with the given parameters was implausible, and although we cannot verify that the treatment-dependent error model is correct or that it has been parameterized perfectly, the example illustrated that the error correction is sensitive to model choice. And second, the fact that the observed counts are perhaps not a surrogate for the true counts means that the treatment effect $\exp(\hat{\beta}_x) = 0.75$, as estimated in Calverley et al. (2007), was not necessarily conservative, as its bias could have been in any direction. When using the treatment-dependent error model, we anyways observed the well-known reverse attenuation effect here ($i.\,e.$ the estimate is $\exp(\hat{\beta}_x) = 0.62$, indicating that 0.75 could be an attenuated version of it). We again refer to simulation 3 of Section 3 for an example that, however, induced reverse attenuation.

Generally, the complexity of error models increases the requirements on the validation or expert data, as has also been discussed by Muff and Keller (2015), where non-constant

error variances and non-constant variances of latent variables were considered. Here, prior information about misclassification in the response of the TORCH trial was extracted from the data presented in Frei et al. (2015), an observational study that aimed to assess the mis-counting error in patient's self-reported numbers of exacerbations. An implicit assumption in the context of prior information transfer typically is that this external validation data are transportable to the study of interest, *i. e.* that the circumstances under which the valida-tion and the study data were collected are comparable, as the information in the validation data does otherwise not lead to sensible priors for the study data analysis. In our example, the validation and study data were collected over the same duration (three years), and the reported exacerbation numbers were comparable. However, it can be objected that the pa-tients in the validation study were not randomly assigned to treatments, which is in contrast to the TORCH study. Moreover, in order to estimate treatment-dependent error model pa-rameters, the validation data were split into two subsets, one for the patients that received a pharmacotherapy, and one for those that did not. Again, the fact that the treatment-free patients were not a random subset of the study population could lead to biased estimates of error model parameters.

The available information from the validation data naturally influences the uncertainty in the prior information of the error model parameters. It is, however, important to under-stand the mechanisms of the hierarchical model in order to avoid biases that emerge due to nonidentifability problems. We have illustrated that, while it can be safe to use priors with covariance matrices estimated from validation data when the error is a surrogate, the same practice is not recommended in the covariate-dependent case, because the $\gamma^{(0)}$ and $\gamma^{(1)}$ parameters are then confounded with the effect of that respective covariate.

In conclusion, we have discussed that error in count outcomes may bias parameter esti-mates of negative binomial regression models, and that the bias may be in any direction. We have introduced a miscounting error modelling framework, and have highlighted the im-portance and also some difficulties of error modelling, particularly in the context of RCTs, where a crucial assumption is that effect estimates originate from conservative estimation procedures. Probably the best way to circumvent expensive and tricky error modelling pro-cedures is to directly optimize the quality of the data. In the example of the TORCH study this could have been achieved by replacing patient self-reports by ascertained values obtained from an adjudication committee. A possible positive side-effect of such an optimization is that smaller study sampling sizes might be sufficient.

Acknowledgements

of Zurich.

References

Abbrevaya, J. and J. A. Hausman (2004). Response error in a transformation model with application to earnings-equation estimation. *Econometrics Journal 7*, 366–388.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association 45*, 164–180.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and roubstness. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 143*(4), 383–430.

Buonaccorsi, J. P. (1991). Measurement error, linear calibration and inferences for means. *Computational Statistics and Data Analysis 11*, 239 – 257.

Buonaccorsi, J. P. (1996). Measurement error in the response in the general linear model. *Journal of the American Statistical Association 91*, 633 – 642.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: CRC Press.

Buonaccorsi, J. P. and T. Tosteson (1993). Correcting for nonlinear measurement error in the dependent variable in the general linear model. *Communications in Statistics, Theory & Methods 22*, 2687 – 2702.

Calverley, P. M., J. A. Anderson, B. Celli, G. T. Ferguson, C. Jenkins, P. W. Jones, J. C. Yates, and J. Vestbo (2007). Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *New England Journal of Medicine 356*(8), 775–789.

Carroll, R. J. and D. Ruppert (1988). *Transformation and Weighting in Regression*. London: Chapman & Hall.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.). Boca Raton: Chapman & Hall.

Copas, J. B. (1988). Binary regression models for contaminated data (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 50*, 225–265.

Ekholm, A. and J. Palmgren (1987). Correction for misclassification using doubly sampled data. *Journal of Official Statistics 3*, 419–429.

Frei, A., L. Siebeling, C. Wolters, P. Muggensturm, A. Strassmann, M. Zoller, G. ter Riet, and M. A. Puhan (2015). Accuracy of different methods to ascertain COPD exacerbations and implications for treatment effect estimates and sample size. Technical Report.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton: Chapman & Hall/CRC.

Gustafson, P. (2005). On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables. *Statistical Science 20*(2), 111–140.

Held, L., D. Sabanés Bové, and I. Gravestock (2015). Approximate Bayesian model selection with the deviance statistic. *Statistical Science*. Accepted for publication.

Hilbe, J. M. (2011). *Negative Binomial Regression*. New York: Cambridge.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association 103*(410–423), 13–22.

Magder, L. S. and J. P. Hughes (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology 146*(2), 195–203.

Muff, S. and L. F. Keller (2015). Reverse attenuation in interaction terms due to covariate error. *Biometrical Journal*. DOI: 10.1002/bimj.201400157.

Muff, S., A. Riebler, L. Held, H. Rue, and P. Saner (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 64*(2), 231–252.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika 86*, 843–855.

Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics 58*, 675–683.

Paulino, C. D., P. Soares, and J. Neuhaus (2003). Binomial regression with misclassification. *Biometrics 59*, 670–675.

Pearson, K. (1902). On the mathematical theory of errors of judgement. *Philosophical Transactions of the Royal Society of London A 198*, 235–299.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine 8*, 431–440.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 71*, 319–392.

Suissa, S. (2006). Statistical treatment of exacerbations in therapeutic trials of chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine 173*(8), 842–846.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics 8*(2), 158–183.

Wald, A. (1940). The fitting of straigth lines if both variables are subject to error. *Annals of Mathematical Statistics 11*, 284–300.

White, G. C. and R. E. Bennets (1996). Analysis of frequency count data using the negative binomial distribution. *Ecology 77*(8), 2549–2557.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Berlin Heidelberg: Springer.
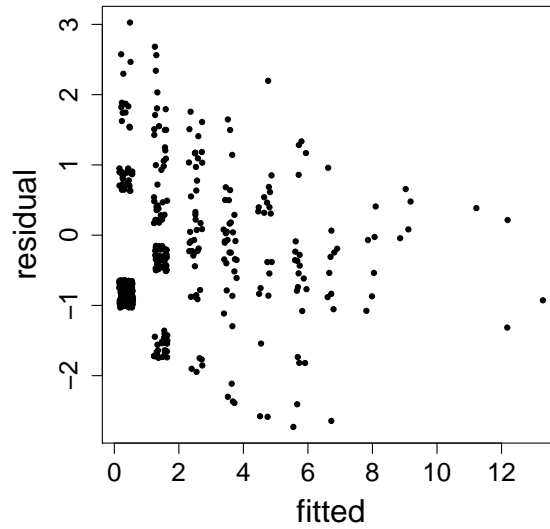
6 Appendix A

Figure 3: Deviance residuals versus fitted values of the negative binomial regression, where the centrally adjudicated exacerbation counts from the validation study were fitted against the respective patient self reports. A small jitter has been added to both the fitted values and the residuals.

7 Appendix B

Let us look at the hierarchical model (8) with log link for the regression model and identity link for the error model. The rate ratio of the treatment effect is then given as

$$\frac{\exp(\beta_0 + \beta_x)}{\exp(\beta_0)} = \exp(\beta_x) \ ,$$

whereas the expected value of the naive estimate is

$$\frac{\gamma_0 + \gamma_1 \cdot \exp(\beta_0 + \beta_x)}{\gamma_0 + \gamma_1 \cdot \exp(\beta_0)} \ .$$

To show that the naive estimate is always attenuated, *i. e.* biased towards 1, distinguish two cases.

Case 1: $\beta_x < 0$
Then $\exp(\beta_x) < 1$, and therefore

$$\gamma_0 > \gamma_0 \exp(\beta_x) \ .$$

Adding $\gamma_1 \exp(\beta_0 + \beta_x)$ on both sides gives

$$\gamma_0 + \gamma_1 \exp(\beta_0 + \beta_x) > \underbrace{\gamma_0 \exp(\beta_x) + \gamma_1 \exp(\beta_0 + \beta_x)}_{=\exp(\beta_x)\cdot(\gamma_0 + \gamma_1 \exp(\beta_0))} \ .$$

Finally,

$$\frac{\gamma_0 + \gamma_1 \cdot \exp(\beta_0 + \beta_x)}{\gamma_0 + \gamma_1 \cdot \exp(\beta_0)} > \exp(\beta_x) \ .$$

which shows that the naive estimate is biased upwards. Moreover, the naive estimate is bounded by 1, which shows that it is biased towards 1.

Case 2: $\beta_x > 0$

In this case, $\exp(\beta_x) > 1$, and exactly inverted arguments as in case 1 show that the naive estimate then lies between 1 and $\exp(\beta_x)$, *i. e.* bias towards 1 and thus attenuation of the rate ratio.