Biostatistics, 2024, **25**, 904–918
https://doi.org/10.1093/biostatistics/kxad027
Advance Access publication September 16, 2023

A Bayesian nonparametric approach to correct for underreporting in count data

Serena Arima[1,*], Silvia Polettini[2], Giuseppe Pasculli[3], Loreto Gesualdo[4], Francesco Pesce[5], Deni-Aldo Procaccini[6]

[1]Department of Human and Social Sciences, University of Salento, Via di Valesio, 73100, LECCE, Italy
[2]Department of Social and Economic Sciences, Sapienza University, P.le Aldo Moro, 5, 00185 ROMA, Italy
[3]Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University, Via Ariosto, 25, 00185 Roma RM, Italy
[4]Section of Nephrology, Department of Precision and Regenerative Medicine and Ionian Area (DiMePre-J), Azienda Ospedaliero Universitaria Consorziale Policlinico di Bari, Piazza Giulio Cesare, 11 - 70124 Bari, Italy
[5]Division of Renal Medicine, "Fatebenefratelli Isola Tiberina–Gemelli Isola", 00186 Rome, Italy
[6]Section of Nephrology, Department of Precision and Regenerative Medicine and Ionian Area (DiMePre-J), Azienda Ospedaliero Universitaria Consorziale Policlinico di Bari, Piazza Giulio Cesare, 11 - 70124 Bari, Italy

[*]To whom correspondence should be addressed: Serena Arima. Email: serena.arima@unisalento.it

SUMMARY

We propose a nonparametric compound Poisson model for underreported count data that introduces a latent clustering structure for the reporting probabilities. The latter are estimated with the model's parameters based on experts' opinion and exploiting a proxy for the reporting process. The proposed model is used to estimate the prevalence of chronic kidney disease in Apulia, Italy, based on a unique statistical database covering information on $m = 258$ municipalities obtained by integrating multisource register information. Accurate prevalence estimates are needed for monitoring, surveillance, and management purposes; yet, counts are deemed to be considerably underreported, especially in some areas of Apulia, one of the most deprived and heterogeneous regions in Italy. Our results agree with previous findings and highlight interesting geographical patterns of the disease. We compare our model to existing approaches in the literature using simulated as well as real data on early neonatal mortality risk in Brazil, described in previous research: the proposed approach proves to be accurate and particularly suitable when partial information about data quality is available.

KEYWORDS: Chronic kidney disease (CKD); Compound Poisson distribution; Data quality; Dependent Dirichlet process; MCMC; underreporting.

1. INTRODUCTION AND MOTIVATION

Data quality underpins the effective use of information in all data driven processes. A quality issue typical of surveillance, notification, and other official registers is underreporting of events. Failing to account for the missing cases may lead to severe underestimation of vital statistics, affecting incidence and prevalence of diseases, and morbidity and mortality rates in turn, thus jeopardizing the aims of a National Health System (NHS). Defective reporting may either occur because not all cases seek healthcare (underascertainment), or due to the inadequacy to detect symptomatic cases that have sought medical advice (see Gibbons *and others*, 2014).

Received: June 16, 2022. **Revised:** June 6, 2023. **Accepted:** August 21, 2023

Consider the number of events of interest over a set of m areas, denoted by T_i, with

$$T_i \sim \text{Poisson}(E_i\theta_i), \qquad i = 1, \ldots, m \tag{1.1}$$

where θ_i is the event rate, and E_i is the known number of exposed in the ith area. The rates θ_i are usually related to a set of covariates $X_1, \ldots, X_p$; a wide variety of regression models can be specified, possibly also allowing for spatio-temporal variation, depending on the application at hand. To estimate the event intensities θ_i, $i = 1, \ldots m$ and predict the true counts T_i $i = 1, \ldots m$ based on underreported observations $Y_i \leq T_i$, $i = 1, \ldots m$, all the available information must be exploited. The Poisson model has been extended to account for underreporting in various ways, all relying on expert information and/or auxiliary data on the reporting process: Bailey *and others* (2005) extended the censored Poisson regression model in Caudill and Mixon (1995) assuming the counts of underreported areas are the lower bound for the true nonobserved counts. This approach requires precise information about which areas are underreported. Bailey *and others* (2005) proposed *ad hoc* procedures to partition the areas into observed or censored. As commented in Stoner *and others* (2019, formula (3) and subs.), the approach based on censored likelihoods cannot allow for the severity of the underreporting, which limits the quality of the prediction of the true counts. Moreover, to estimate the underlying counts at least some of the units must be completely observed. An alternative approach, modeling the true count prevalence as well as the probability of underreporting, relies on Poisson stopped-sum distributions (see Johnson *and others*, 2005, Section 4.11). A vast literature (see e.g., Whittemore and Gong, 1991; Winkelmann, 1996; Stoner *and others*, 2019; Dvorzak and Wagner, 2016; Li *and others*, 2003, to mention but a few) follows this approach, often working with the compound Poisson model (CPM) described in the next section. In particular, Stoner *and others* (2019) model the reporting probability hierarchically through a logistic regression using suitable covariates. In assessing early neonatal mortality in Minas Gerais, Brazil, de Oliveira *and others* (2022) also refer to the compound Poisson model. Instead of specifying a regression model for the reporting probabilities, their approach relies on a data quality variable, that allows to cluster the areas and identify the ones that are nearly unaffected by underreporting. Following a similar approach, in this article, we propose a Bayesian nonparametric model for dealing with underreported count data when prior information about data quality is available but not completely accurate. Compared to the previous literature, our proposal requires milder information about data quality since clusters are not set *a priori*, but inferred at the model fitting stage based on a proxy for underreporting.

As an application we consider registry data on chronic kidney disease (CKD) in Apulia, which is one of the most deprived regions in Italy. As with all chronic diseases, CKD has a strong social (van Oostrom *and others*, 2016) as well as economic impact (Pontoriero *and others*, 2007). The "awareness gap" and the under-recognition of CKD in primary care settings have been well documented (Pesce *and others*, 2022). Accurate estimation of the prevalence of CKD would support policy makers in optimizing fund allocation. However, as clearly explained in the seminal paper by Hart (1971), availability of medical care tends to vary inversely with the need for it in the population. Consequently, experts suspect substantial underreporting in those areas that lack medical care the most. Moreover, Apulia is a very heterogeneous region from both a geographical and a sociocultural perspective, and we may expect different patterns of underreporting across the region. In this respect, to identify a proxy for data quality, we build on Fitzpatrick *and others* (2004) and Lin *and others* (2015), who indicate economic and physical barriers like costs or inadequate transport infrastructures as the factors most often reported to affect one's ability to reach specialized premises of the regional health system.

The article is organized as follows: Section 2 reviews the compound Poisson model. Section 3 describes the proposed approach; in Section 4 using a simulation study and the data in de Oliveira *and others* (2022), our proposal is compared with competing models. In Section 5, we discuss the application of our approach to CKD data. We conclude with a short discussion in Section 6.

2. THE COMPOUND POISSON MODEL

Underreporting of counts may be accounted for by assuming that in area i the T_i events in (1.1) get reported independently, each with probability ϵ_i $(i = 1, \ldots, m)$. For the observed counts Y_i, it therefore holds that $Y_i|T_i, \epsilon_i \sim Bin(T_i, \epsilon_i)$. Combining the above model with (1.1) and marginalizing over T_i leads to

$$Y_i|\theta_i, \epsilon_i \sim \text{Poisson}(E_i\theta_i\epsilon_i), \tag{2.2}$$

i.e., the CPM. In this framework, Winkelmann and Zimmermann (1993) propose modeling the reporting probabilities ϵ_i through covariates via a logistic regression, the so called Poisson-logistic ("Pogit") model. See also Dvorzak and Wagner (2016) and Stoner *and others* (2019) in an epidemiological context. de Oliveira *and others* (2022) assume instead that areas can be *a priori* grouped, with clusters arranged in decreasing order of data quality. That is, they assume a cluster-specific structure on $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)$, with $\epsilon_i = 1 - h_i^T\boldsymbol{\gamma}$, $i = 1, \ldots, m$, where $h_i = (h_{1i}, \ldots, h_{Ki})^T$ is the cluster membership vector defined according to the following split-coding scheme: if area i belongs to cluster j, then $h_{li} = 1$ for all $l \leq j$, otherwise $h_{li} = 0$. They also assume $\boldsymbol{\gamma} \in [0, 1)$ and $\sum_{j=1}^{K} \gamma_j < 1$ to ensure nonzero mean for the associated Poisson distribution. Under model (2.2), an issue of identifiability arises in the absence of any completely reported observations: indeed, whereas the product $\theta_i\epsilon_i$ is identified from the observations, θ_i and ϵ_i are not. When validation data are available, they can be used to resolve the identifiability issue. Otherwise, constraining the parameters based on expert knowledge is necessary. A thorough discussion of the identifiability issue is reported in Papadopoulos and Santos Silva (2012). de Oliveira *and others* (2022) also discuss the different approaches followed in the literature to inform the reporting process within the CPM. Among those, the proposal in Stoner *and others* (2019) is appealing because it only requires an informative prior for the mean reporting rate to be specified. However, as observed by de Oliveira *and others* (2022), trustful information on the overall reporting process is difficult to obtain, whereas partial prior knowledge arising from local inventory studies or experts' opinion may be more readily available. Given a partitioning of areas in decreasing order of the reporting probability, they show that the imposed clustering implies a reduction in the parametric space under which identification is guaranteed by specifying an informative prior for the reporting probability in the cluster where this probability is highest. We observe that even when a good proxy for data quality is available, experts' opinion is still crucial in partitioning the vector of the reporting probabilities $\boldsymbol{\epsilon}$ into a suitable number K of different groups. Moreover, the proxy would usually be an imprecise measure of data quality in practice, with a subsequent imprecision in cluster membership. The real data application described in de Oliveira *and others* (2022) indeed shows a certain sensitivity of parameter estimates to the number of clusters and to cluster assignment. Moreover, the approach just mentioned does not take into account the uncertainty associated with the given partition.

3. THE PROPOSED APPROACH

To tackle the above mentioned limitations, we propose modeling the clustering of the areas jointly with the other model components by adopting a Bayesian nonparametric approach under which the partitioning of the areas is driven by a variable that is a proxy for data quality. To each cluster, we attach a reporting probability that is estimated jointly with the other parameters of the model. Covariates are used to inform both the true count-generating process and the reporting process. This enables us to account for underreporting in estimating prevalences, at the same time deriving a partition of areas according to the probability of underreporting. The clustering of areas is not performed in advance: it is obtained by specifying a nonparametric prior on the reporting probabilities, specifically a generalization of the Dirichlet Process, that allows us to make the clustering of the areas depend on a proxy for data quality. This modeling choice amounts to giving areas with similar covariate values higher prior probabilities of belonging to the same cluster. We model underreported counts through the compound Poisson distribution (2.2) with log-linear

modeling of the rates, supplemented by a nonparametric assumption for the distribution of the reporting probabilities. Working with standardized covariates, in the sequel we omit the intercept β_0. The proposed model reads as follows:

$$Y_i|\theta_i, \epsilon_i \sim \text{Poisson}(E_i\theta_i\epsilon_i), \quad i = 1, \ldots, m \tag{3.3a}$$

$$\log(\theta_i) = \beta_1 X_{1i} + ... + \beta_p X_{pi} + u_i + s_i \tag{3.3b}$$

$$u_i|\sigma_u^2 \sim N(0, \sigma_u^2) \tag{3.3c}$$

$$s_i|\sigma_s^2 \sim \text{ICAR}(\sigma_s^2) \tag{3.3d}$$

$$\epsilon_i|z, G_z \sim G_z, \tag{3.3e}$$

where ICAR denotes the intrinsic conditional autoregressive distribution of Besag *and others* (1991) with variance parameter σ_s^2, and G_z is a realization the aforementioned Bayesian nonparametric prior described in the next section. The random effects were introduced to account for lack of fit and capture extra variability, spatially and non spatially structured. We adopted the Poisson distribution since it is ubiquitous in epidemiological studies, especially thanks to the wide use of the Besag–York–Mollié model and its BYM2 extension for estimating relative risks in disease mapping. However, as suggested by a referee, the negative binomial may be a valuable alternative to model overdispersion and reducing autocorrelation in the MCMC sampling. In our proposal, we keep the linear predictor in equations (3.3b)–(3.3d) as simple as possible. However, any other parametrization could be specified (e.g., splines or ridge regression, spatio-temporal effects, ...), to increase the model's flexibility and fit.

3.1. Prior specification

3.1.1. *Reporting probabilities*

In a Bayesian framework, nonparametric priors like the Dirichlet process (DP, Ferguson, 1973) and its extensions (Pitman and Yor, 1997; Ishwaran and James, 2001) are recognized as effective tools to improve model flexibility, with a wide range of applications (see Müller *and others*, 2015, for a review). By a modification of Sethuraman's representation of the DP, MacEachern (1999) introduced the Dependent Dirichlet process (DDP). His construction generalizes the DP to specify nonparametric distributions whose realizations are dependent, with the degree of dependence governed by the level of a covariate, and DP-distributed for every possible predictor value z (Quintana *and others*, 2022). Recent related literature has drawn its attention to the definition of predictor-dependent stick-breaking priors that retain some of the key characteristics of the Dirichlet process, most notably, the clustering of units, and that can be made dependent on a covariate. We refer to Quintana *and others* (2022) for an analytic description of the wide range of approaches in this context.

Within the wide class of predictor-dependent stick-breaking priors, we rely on the probit stick-breaking process (PSBP) of Chung and Dunson (2009) (see also Rodriguez and Dunson, 2011), under which the mixing weights arise through a probit model on the covariate. More precisely we specify G_z in equation (3.3e) as:

$$G_z(\cdot) = \sum_{j=1}^{\infty} \left(\Phi(\eta_j(z)) \prod_{l<j} (1 - \Phi(\eta_l(z))) \right) \delta_{\psi_j}(\cdot), \tag{3.4}$$

where the weights are obtained through normally distributed random variables $\eta_j(z) = z'\gamma_j$ transformed to the unit interval via the standard normal cdf (see Chung and Dunson, 2009; Quintana *and others*, 2022), whereas $\delta_{\psi_j}(\cdot)$ denotes the Dirac measure at location ψ_j, and ψ_j $(j = 1, 2, \ldots)$ are independent draws from G_0, that do not depend on the covariates. In our model (3.3), we assume a PSBP for G_z with weights depending on a proxy for data quality. In model

fitting, as common practice in the literature, the stick-breaking representation (3.4) is truncated to a large value K'.

Following Rodríguez *and others* (2010), the sequence of atoms ψ_j is constructed by introducing a partial ordering on a sample from a baseline measure G_0 using the following mechanism: ψ_1 is drawn from $G_0^* = U[a_0, b_0]$ and $\psi_j \sim G_{0j}^*$, $j = 2, \ldots$ with $G_{0j}^* = U[a_1, b_1]$ with $a_1 < b_1 < a_0$ to ensure $\psi_j < \psi_1$ for $j = 2, \ldots$. The choice of a_0 and b_0 should reflect the prior information about the best reporting probability. In the applications described in Section 5 and Section 3 of the supplementary material available at *Biostatistics* online, we specified $G_0^* = U[0.9, 1]$ and $\psi_j \sim G_{0j}^*$, $j = 2, \ldots$ with $G_{0j}^* = U[0.3, 0.9]$ for $j > 1$.

Rodríguez *and others* (2010) comment that the use of order constraints allows generating skewed distributions. This behavior is particularly suitable in our applications where the reporting probabilities are expected to be right-skewed, to reflect good data quality for some areas and entails identification of the model. Concerning this latter aspect, we rely on the results in de Oliveira *and others* (2022), who show that under a clusterization of the areas, eliciting a prior on the reporting probability in the best reported areas using experts' knowledge is sufficient for identification. We build on the above result, while imposing the above mentioned partial ordering at the level of the hyperparameters of the nonparametric prior, ψ_j.

The reporting probabilities ϵ_i, $i = 1, \ldots, m$ therefore follow a distribution that is indexed by z. Considering the PSBP as a random partition model (Müller and Quintana, 2010), this effectively makes the expected counts to be clustered in a way that is driven by the covariate associated with the reporting accuracy: units with similar covariate values have higher probability to be clustered together. Such a feature is particularly important in a context where the auxiliary variable is just a proxy and not a completely accurate determinant of the reporting probability. In the PSBP specification (3.4), we assume $\gamma_j \sim N(\mu_\gamma, \sigma_\gamma^2)$, with standard Normal hyperprior, $\mu_\gamma \sim N(0, 1)$, and $\sigma_\gamma^2 = 1$.

3.1.2. *Model parameters*

For the area random effect $u_i \sim N(0, \sigma_u^2)$ in (3.3c), we specify $\sigma_u^{-2} \sim \text{Gamma}(a_u, b_u)$, independently, $i = 1, \ldots, m$. For the variance parameter σ_s^2 of the ICAR spatial random effect, we assume $\sigma_s^{-2} \sim \text{Gamma}(a_s, b_s)$. Furthermore, we set $\beta_j \sim N(0, \sigma = 10^2)$, $j = 1, \ldots, p$ in the regression component; for the random components' priors, the hyperparameters a_u, b_u, a_s, b_s have all been fixed to 0.01.

3.2. Posterior computations

Since the posterior distributions are not available in closed form, we draw samples from the posterior distributions of model parameters using Markov Chain Monte Carlo (MCMC) techniques. The model has been estimated using NIMBLE (de Valpine *and others*, 2017). The code used for fitting the proposed model can be found in the supplementary material available at *Biostatistics* online and at `https://github.com/serena-arima/Biostat-SuppMaterials`, along with synthetic data.

Besides inferring the event rates θ_is, one may be also interested in the prevalences $T_i, i = 1, \ldots, n$. As in Stoner *and others* (2019, see eq. (9)), we can draw samples from the posterior predictive distribution of $T_i|\mathbf{y}$ obtained by integrating the full conditional distribution with respect to the joint posterior of the model parameters. To this aim, notice that under the compound Poisson model, the conditional distribution of $T_i - Y_i|\Omega, \mathbf{y}$ is again Poisson $((1 - \epsilon_i)E_i\theta_i)$, Ω representing the vector of all model parameters.

4. MODEL ASSESSMENT

In this section, we investigate the performance of the proposed approach in a controlled setting. Model performance is assessed by evaluating the predictive ability as well as the overall fit under

Table 1. wAIC of all models and each simulation scenario averaged over the $D = 50$ generated data sets.

	M_{Ref}	M_{Pois}	$M_{\text{deOliveira}}$	M_{Prop}
S_1	762.83	767.10	763.80	762.54
S_2	727.35	731.07	729.61	728.39
S_3	737.29	745.28	743.67	741.55

different patterns of the reporting probabilities. We also compare our results with those obtained by de Oliveira *and others* (2022) using data on early neonatal mortality risk in Brazil. All the details are reported in the supplementary material available at *Biostatistics* online.

For the simulation study, we generate $D = 50$ data sets, comprising $m = 75$ observations drawn from model (3.3a)–(3.3c). We assume two covariates x_1, x_2 are available for the intensity, and generate their values from independent standard normal distributions. We fix $\beta = (0.1, -0.5)$ and draw the responses $y_i, i = 1, \ldots, n$ from Poisson r.v.'s with means $E_i\theta_i\epsilon_i$, with $\theta_i = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + u_i)$, $u_i \sim N(0, 1)$ and $E_i \sim \text{Unif}(15, 2000)$. The reporting probabilities ϵ_i are categorized into $K = 4$ levels and specified according to three different scenarios: S_1: $\boldsymbol{\epsilon} = (1, 0.95, 0.90, 0.85)$; S_2: $\boldsymbol{\epsilon} = (0.98, 0.95, 0.90, 0.80)$; and S_3: $\boldsymbol{\epsilon} = (0.90, 0.80, 0.70, 0.60)$. The levels of $\boldsymbol{\epsilon}$ are inversely related to the values of the auxiliary data quality variable z, which is generated as a mixture of $K = 4$ normal distributions with $\boldsymbol{\mu} = (-1, 1, 3, 8)$, $\boldsymbol{\sigma^2} = (2, 1, 2, 3)$, and weights $\boldsymbol{w} = (0.3, 0.3, 0.2, 0.2)$, resulting in respectively 22, 23, 22, and 15 observations in each group. Prior distributions and hyperparameters for the proposed model have been specified as in Section 3.1; hyperparameters a_0 and b_0 have been specified to reflect prior information about the highest reporting probability.

The following models are compared:

- M_{ref}: the model in de Oliveira *and others* (2022) with the true number of clusters ($K = 4$) and the group membership correctly specified through the vectors $h_i, i = 1, \ldots, m$;
- M_{Pois}: the compound Poisson model not accounting for underreporting;
- $M_{\text{deOliveira}}$: the model in de Oliveira *and others* (2022) with a wrong number of clusters ($K = 6$) and group membership vectors h_i specified according to the sextiles of z
- M_{prop}: the proposed model as specified in Section 3.

In the simulation study, M_{ref} is the benchmark. In practice, the exact definition of the clusters based on a suitable discretization of the proxy variable z is subjective. To investigate the effect of alternative definitions of the clustering structure, model $M_{\text{deOliveira}}$ defines $K = 6$ clusters using the sextiles of the distribution of z. For each simulation, we run 2 chains and allow for 80 000 iterations, with a burn-in of 5000 and a thinning rate of 5, resulting in $N = 5000$ draws from each posterior distribution. For each model, we compute the Watanabe–Akaike information criterion (wAIC) originally proposed in Watanabe (2010): the model with the smallest wAIC is to be preferred. Table 1 shows the averaged wAIC over the simulated data sets of each model under the different scenarios. As expected, M_{ref} has the lowest wAIC in all simulation scenarios, followed closely by the proposed model. In S_1 and S_2, where the reporting probabilities are large, all models perform very similarly, while in S_3 a greater difference in terms of overall fit is detected between the proposed model and the competing ones. In terms of wAIC, the proposed model is always close to the reference model even in S_3, where the underreporting is substantial.

We also compare models in terms of the relative mean squared error (RMSE) of the estimates of the event rates θ_i under each simulation scenario. Let $\widehat{\theta}^M_{ij(d)}$ be the jth draw from the posterior distribution of θ_i under model M for the d-th data set. The RMSE of the estimator of θ_i for model M and unit $i = 1, \ldots, m$ is evaluated as

$$\text{RMSE}_i^M = \sqrt{\frac{1}{D \cdot N} \sum_{d=1}^{D} \sum_{j=1}^{N} \frac{(\widehat{\theta}^M_{ij(d)} - \theta_i)^2}{\theta_i}}. \tag{4.5}$$

As shown in Figure 1 (values on a log-scale), the reference model has the best performance in all scenarios, with the Poisson having the poorest behavior. Under S_1, the RMSEs of the proposed models and its competitor $M_{\text{deOliveira}}$ are very similar. This is not surprising, since a large proportion of the areas have high reporting probabilities ($\epsilon_i \geq 0.9$ for 90% of the units). However, when the reporting probabilities decrease, as in S_2 and S_3, the proposed model M_{prop} estimates θ_i more accurately than the competing models. Notably, a different though coherent choice of the clustering structure results in progressively deteriorating performances of model $M_{\text{deOliveira}}$. As expected, all models severely deviate from the reference model, with the proposed model still improving over the alternative ones. When highly reliable prior information about the clustering structure is not available, our proposal seems to achieve reasonable and robust results.

To further investigate the robustness of the proposed model when the best quality cluster has a relatively small number of observations, we repeat the same simulation scheme, but set the cluster weights to $\boldsymbol{w} = (0.05, 0.15, 0.20, 0.6)$, resulting in respectively 4, 11, 15, and 45 observations in each group. Figures 1, 2, and 3 in the supplementary material available at *Biostatistics* online show that the competing models are not able to produce accurate estimates, while the proposed one seems to be robust in all scenarios.

As suggested by a referee, we also investigate a setup in which all areas have the same reporting probability ($K = 1$). We simulate data as in the previous schemes but set the common reporting probability to $1, 0.9, 0.8, 0.7$, respectively, leading to four different scenarios. Here, the values of the proxy for data quality z are generated from a standard normal distribution. Figure 4 in the supplementary material available at *Biostatistics* online shows the RMSEs (on the log scale) of $\widehat{\theta_i}$ under the proposed model (x-axis) and under the Poisson model (y-axis) for each scenario. As expected, in the absence of underreporting ($\epsilon = 1$), the Poisson model tends to perform better than the proposed one. However, as the underreporting probability increases, the Poisson model is less satisfactory and the corresponding RMSEs are significantly higher than those obtained with the proposed model for all areas.

All the approaches in the literature rely on external information to account for the underreporting; likewise, the role of the proxy variable is crucial in our proposal. However, the appropriateness of the proxy cannot be checked and one should solely rely on experts' opinion. To investigate this aspect, we performed a simulation experiment where we weakened the proxy so that the implied clustering structure increasingly differs from the actual reporting process. Figure 5 in the supplementary material available at *Biostatistics* online compares the model using the original and the corrupted proxy, respectively, and shows that our proposal has a certain robustness, provided the proxy is at least moderately correlated with the original one. In Section 5 of the supplementary material available at *Biostatistics* online, we assess our proposal and compare it to the model in de Oliveira *and others* (2022) based on early neonatal mortality risk data in their paper. Substantial agreement in terms of covariate effects, clustering structure, and risk ratios is found.

5. UNDERREPORTING OF CKD IN APULIA

CKD is defined by a gradual loss of kidney function over a period of months or years. At the early stage of the disease there are typically no symptoms, whereas at later stages it might have severe complications such as heart disease, high blood pressure, bone disease, or anemia. The decrease in renal functionality is measured according to the glomerular filtration rate (GFR) (moderate, severe, or end-stage renal disease). Very low GFR is an indication for renal replacement therapies (i.e., dialysis or transplantation). It is well known that the earlier a patient's access to nephrological care, the better the clinical, economic, and psychological outcome (Smart *and others*, 2014). To achieve this goal, which is not only clinical but also a public health and economic objective, we need to recall the construct of patient's frailty as the "inefficiency (failure) of a complex system", e.g. an outcome resulting from the possible multiple interactions between clinical, functional, socio-economic, and environmental determinants. Associations between the risk of CKD and socio-economic factors such as poverty, cultural level, familiar status, and social deprivation have been

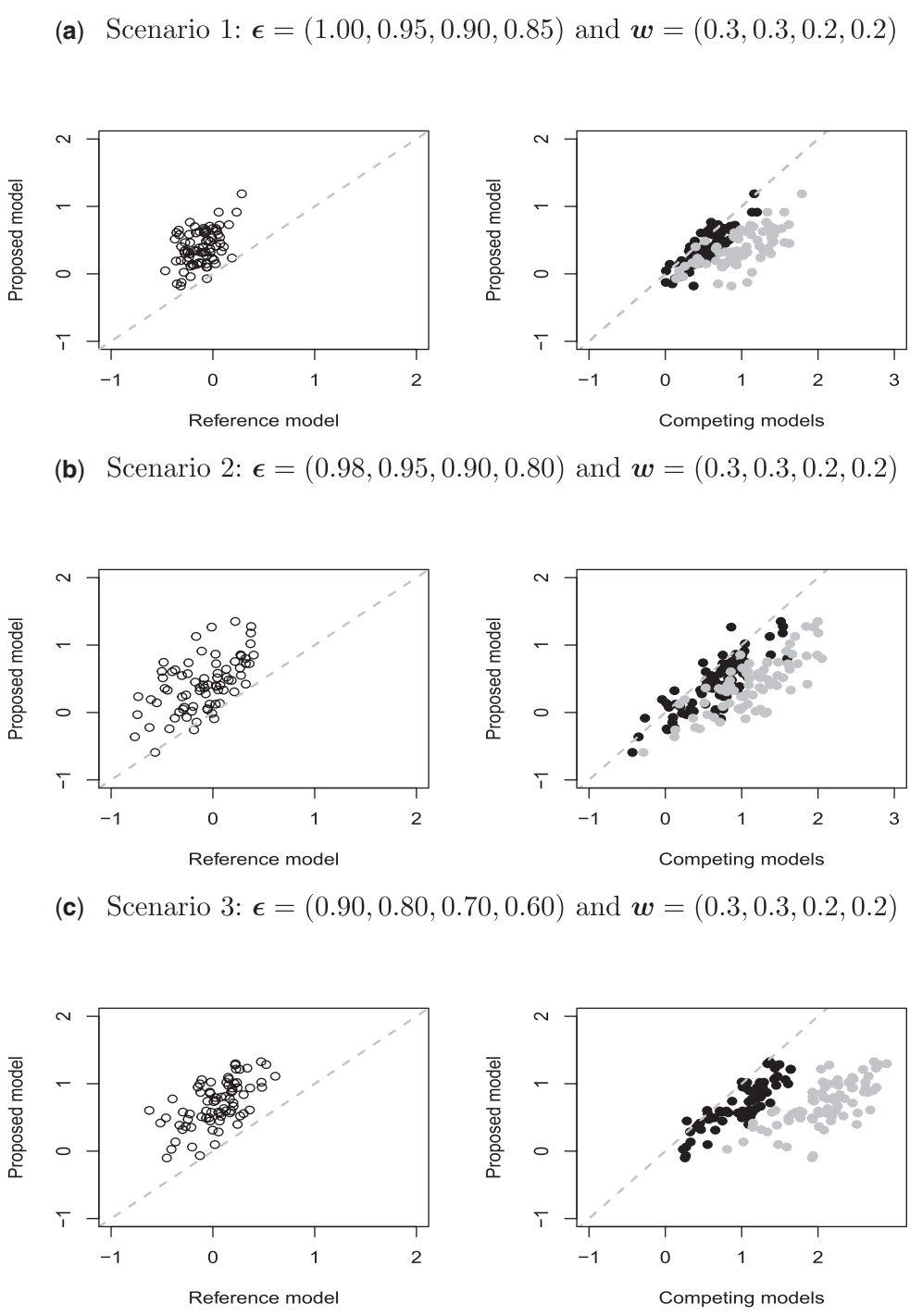

Figure 1. Results of the simulation study. Each line corresponds to a specific Scenario, defined by setting the reporting probabilities and their relative weights as reported on top. The left panel in each line show the RMSEs (on log scale) of the estimates of the event rates θ_i obtained with the reference model versus the those obtained with the proposed approach. The right panels report the RMSEs (on log scale) of the estimates obtained with $M_{\text{deOliveira}}$ (black dots) and M_{Pois} (grey dots) versus those obtained with the proposed model.

documented (Shlipak *and others*, 2021; Caskey, 2016). Low socioeconomic status correlates with a combination of clinical risk factors such as diabetes and hypertension; it is also linked to lifestyle behavior, including dietary patterns, some of which have been associated with favorable CKD outcomes (Banerjee *and others*, 2016). Moreover the socio-economic status impacts on the access to appropriate medical care (e.g., renal/peritoneal dialysis, renal transplants: Hossain *and others*, 2009), also through the distance from the closest health facility (Bigogo *and others*, 2010).

CKD also impacts on patient's life and their network: indeed, caregivers are heavily involved in terms of time, quality of life, and economic support. Accurate estimates of CKD prevalence would aid policymakers in allocating funds more efficiently. However, as commented in Section 1, substantial underreporting is deemed to affect the CKD reports in Apulia (Pesce *and others*, 2022). Our goal here is to map the relative risk of CKD in Apulia region and to identify factors that are possibly associated with the event occurrence. To this aim, we refer to a unique statistical database for CKD covering information on $m = 258$ municipalities in Apulia for years 2011–2013, constructed retrospectively with the endorsement of ARES-Puglia (namely, the Regional Health Agency). In compliance with the current privacy legislation (D.lgs. 196/2003, articles 75-94), the archives of years 2011, 2012, and 2013 concerning the consumption of resources (e.g., hospital admissions, outpatient health services, consumption of drugs), ticket exemptions and causes of death of the patients are made available. CKD patients are categorized from Stage 1 to Stage 5 based on the value of renal function, as reported in the SDOs (the Italian acronym for Hospital discharge records). This database provides considerable insight into the structural characteristics of the disease. Besides the total number of CKD patients, that is recorded for each municipality, the database also contains several socioeconomic information provided by the National Institute of Statistics (ISTAT) at the municipality level.

Experts suspect that crude counts are not trustworthy and that in some areas, especially in the province of Foggia, the number of CKD patients is significantly underreported. This is probably to be ascribed to social and organizational hindrances that prevent patients to reach nephrological care-points. In this context, it is deemed that a model accounting for underreporting might improve estimates of CKD rates, thus allowing for an appropriate allocation of funds and healthcare facilities for CKD patients; in particular, stimulating early referral through appropriately targeted policies is essential to mitigate the disease progression and to reduce inequalities between the communities of the same region.

In the application, y_i is the observed CKD cases for each municipality in the Apulia CKD database ($i = 1, \ldots, 258$); the age-sex-specific expected CKD cases E_i have been computed by means of the usual naive estimator $E_i = \sum_k \sum_{j=1} N_{ijk} q_{jk}$, where the specific age and sex rate equals $q_{jk} = \frac{\sum_{i=1}^{258} y_{ijk}}{\sum_{i=1}^{258} n_{ijk}}$ in which y_{ijk}, n_{ijk}, and N_{ijk} are respectively the counts, number of patients at risk, and national total population for the ith area, jth age range (5 years bins from 60 to 80 years), and kth sex (i.e., Male or Female).

We fit the proposed model (3.3) using demographic density (X_1), mean age (X_2), unemployment rate (X_3), and percentage of house ownership (X_4) as covariates affecting the risk of CKD in (3.3b). A spatial effect $s = (s_1, ..., s_m)$ that quantifies the influence of neighboring areas has been included, see (3.3d). Since all covariates have been standardized, the intercept is not estimated. Building on the findings in Fitzpatrick *and others* (2004) and Lin *and others* (2015), difficulties to access healthcare facilities are considered as barriers to detect and notify events to the surveillance system (Jug, 2014). For inferring the distribution of the underreporting probability in (3.3e) we computed a proxy (Z) measuring the minimum travel distance from the center of each area to the closest specialized health facility using the Google Maps distance calculator tool.

For the specification of the prior distributions for the model parameters, we use the same settings described in Sections 3.1.1 and 3.1. Estimates are obtained by MCMC techniques: we allow for 2 chains and 15 5000 iterations (5000 burn-in and thinned over 10). Convergence has been evaluated visually by looking at the trace plots of the regression parameters (see Figure 6 in the supplementary

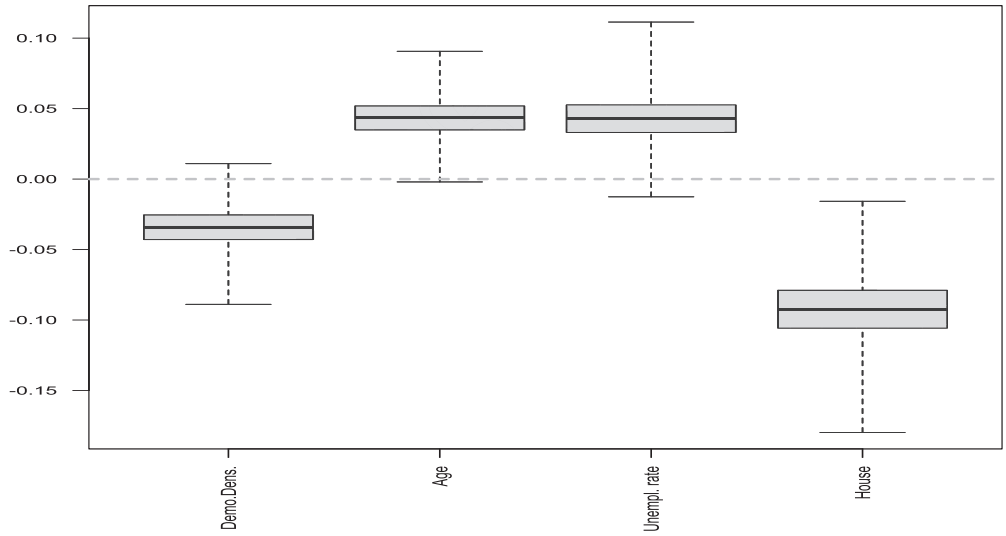

Figure 2. CKD data in Apulia: Posterior distribution of the regression parameters estimated with the proposed model.

material available at *Biostatistics* online). Autocorrelation functions and Geweke's statistics confirm that the MCMC chains reached convergence. As a further validation, we also compute the potential scale reduction factor for all model parameters: all indices are less than 1.03 assuring us about the chains' convergence.

Figure 2 shows the posterior distribution of the regression parameters. Population density seems not to affect the CKD risk (95% credible interval (CI) $[-0.051; 0.001]$) while the average age (CI: $[0.025; 0.079]$) and the unemployment rate have a strong positive impact on the number of CKD cases (CI: $[0.015; 0.069]$). On the other hand, an increase of the percentage of families who are owners of their house produces a substantial decrease of the number of CKD cases (CI: $[-0.121; -0.063]$). These conclusions are in agreement with the medical literature on the field: Mallappallil *and others* (2014) showed that demographic variables such as ageing are significantly associated with an increased risk of CKD, as for other chronic diseases. Moreover, findings in Krop *and others* (1999) based on the National Health and Nutrition Examination Survey (NHANES) show that unemployed people had twice more CKD prevalence than their employed counterparts. On the other hand, the percentage of house ownership can be interpreted as a proxy of the population wealth and, as a consequence, of lifestyles and timely access to appropriate therapies (Chang *and others*, 2020).

Figure 3(b) shows the estimated CKD risks in Apulia. Compared to the observed prevalences reported in Figure 3 (a) (i.e., the ratio of the observed CKD cases over the population in each area), the largest discrepancies can be appreciated in the province of Foggia, where low observed CKD counts correspond to large estimated CKD rates. Since the province is the most deprived area of the whole region, this reflects experts' hypothesis that a combination of social and organizational difficulties hinder patients' access to nephrological care-points. Interestingly, the posterior means of the underreporting probabilities $1 - \epsilon_i$ (Figure 3(c)) are higher in the province and particularly at the borders with other neighboring regions, hence matching with the "inner" and mostly deprived areas (see https://www.agenziacoesione.gov.it/strategia-nazionale-aree-interne/?lang=en). A similar situation can also be elucidated in the Taranto surrounding areas, one of the most isolated and environmental risk exposed area in the Apulia region.

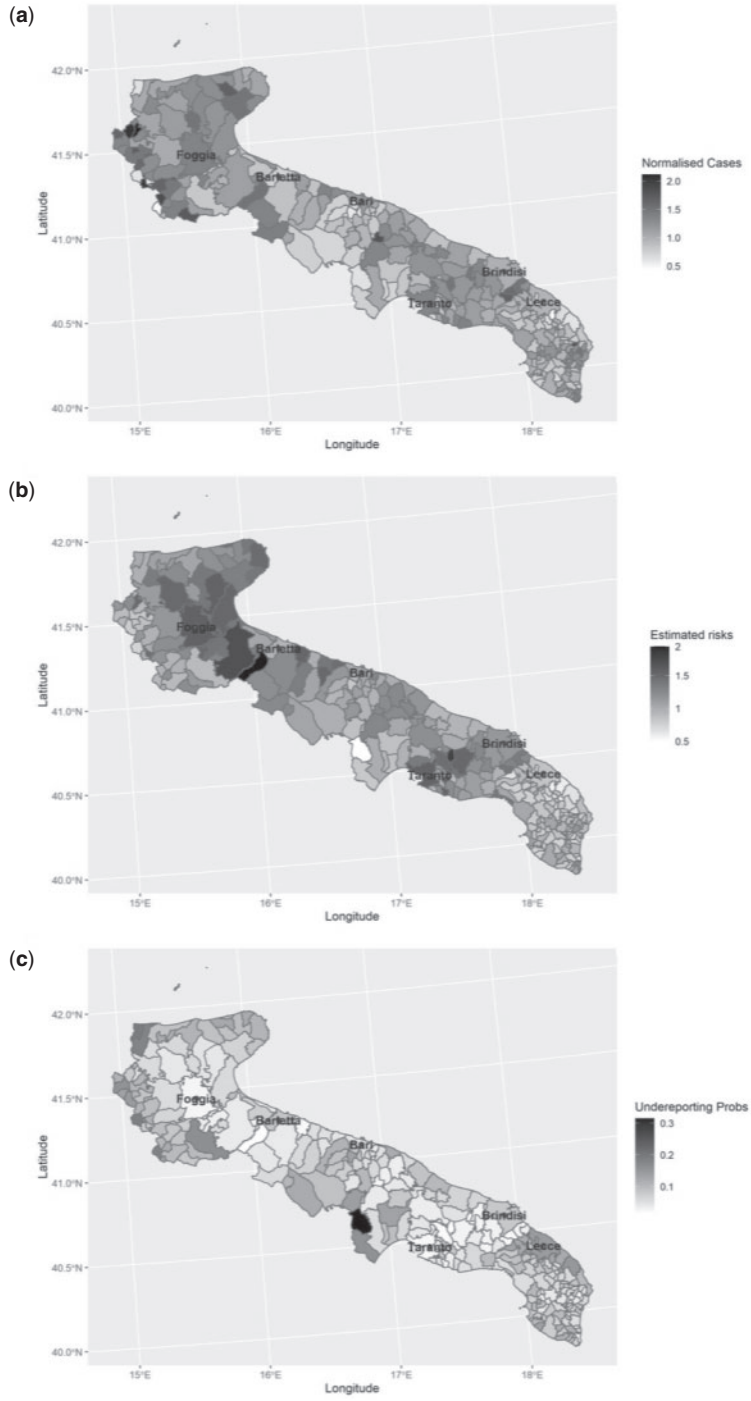

Figure 3. CDK data in Apulia : (a) prevalence (i.e., the ratio of the observed CKD cases and the population in each area). (b) posterior mean under the proposed model for the relative risks θ_i of CKD. (c) posterior mean of the underreporting probabilities $(1 - \epsilon_i)$ under the proposed model.

So far, health policy programs have been targeted to biased figures that do not account for possibly large underreporting. Timely and bias-free estimates of CKD prevalence is clearly essential for accurate and effective planning of any intervention targeted to prevent disease progression and to encourage early referral to the nephrological care points, thus reducing inequalities between the communities of the same region. As a matter of fact, in areas with low quality administrative data, prone to underreporting, patients are more likely to be under-diagnosed due to a reduced number of NHS facilities. Paradoxically, those areas could be considered at low risk of CKD, thus receiving less than needed resources. Our proposed model might provide improved information regarding care needs of the CKD regional population, readily available to policy-makers for public health decisions, hence ensuring appropriate care and affecting those factors that might improve the patient's outcome.

6. DISCUSSION

In this article, we presented a novel Bayesian modeling framework to analyze potentially underreported count data. A semiparametric extension of the compound Poisson model is presented, that relies on covariates used to inform both the intensity and the reporting probability, namely θ_i and ϵ_i, in the absence of validation data sets. Expert's opinion and auxiliary variables expressing data quality throughout the areas are essential and used to inform the model. The proposed method might be a compelling alternative when expert information is available for the reporting rate in the best-reported areas rather than for the average reporting rate across all areas, as in Stoner *and others* (2019), for instance. As in Stoner *and others* (2019) and de Oliveira *and others* (2022), we model the severity of the underreporting, not only its occurrence. While the approach by Stoner *and others* (2019) permits learning the level of underreporting at a fine detail, such learning might be affected by incorrect choice of the regression model. In our proposal, heterogeneity between areas in terms of the reporting process is accounted for by assuming a mixture structure in which different but correlated mixing measures are specified, with the degree of dependence controlled by the covariate values; the parameters $\epsilon_i, i = 1, \ldots, m$ are modelled by a PSBP. Considering the PSBP as a random partition model (Müller and Quintana, 2010), this effectively means the true counts $T_i\ i = 1, \ldots m$ are clustered at the model-fitting stage and in a way that is driven by the reporting accuracy, with units having similar values of the covariate being more likely to be clustered together. This feature is particularly important in a context where the auxiliary variable is just a proxy and not a completely accurate determinant of the reporting probability. This approach can be seen as a nonparametric version of Stoner *and others* (2019), which gives robustness to the procedure. In this respect, we deem it is an improvement over de Oliveira *and others* (2022) where areas are *a priori* grouped and uncertainty of clustering is not accounted for. However, as noticed by one of the referees, the role of the proxy is crucial in all the approaches to underreporting, yet its choice can only be based on experts' opinion and cannot be checked using the observed data. Since in our approach such information enters the model trough the nonparametric prior, we believe our proposal can be considered a flexible alternative, in that it supplies the least amount of prior information about the reporting process. Concerning the model definition, we notice that the regression component can accommodate any specification of the linear predictor for the true Poisson counts and could be extended, for instance, to account for alternative spatial patterns (e.g., putative hazard models taking into account the distance of the municipalities from major polluted sites around the areas).

With respect to the choice of the nonparametric prior, recently, Rigon and Durante (2021) consider the logit stick breaking process (LSBP, see Ren *and others*, 2011) in place of the probit, stressing the ease of interpretation of the logit model in terms of log-odds and the computational advantage of the approach. In fact, whereas PSBP can be approximated by LSBP, and vice versa, the logistic representation encodes a simple generative process for the group labels that may facilitate computations using the results in Polson *and others* (2013).

In the application to CKD data in Apulia, accurate mapping of hazards associated with vital statistics is crucial in targeting health policies that can reduce disease progression and thus affect

mortality. Estimates of event reporting probabilities, which provide a measure of the severity of underreporting, may support in determining the areas where surveillance resources are most needed and effective. Furthermore, we acknowledge that southern Italian regions, like Apulia, experience an important sanitary migration towards the North of Italy (plus many individuals often consult private outpatient specialists at their own expenses, see Toth, 2014). Indeed there are significant organizational variations in how Italian Regions have chosen to carry out their health-delivery tasks, which could lead to disparities in access rates and barriers to health care. This fact, in turn, might definitely impact towards the estimation of the underreporting. We are planning to undertake research over the latter aspect in future works.

SUPPLEMENTARY MATERIAL

Supplementary Material is available online at http://biostatistics.oxfordjournals.org.

CONFLICT OF INTEREST

None declared.

REFERENCES

Bailey, T. C., Carvalho, M. S., Lapa, T. M., Souza, W. V. and Brewer, M. J. (2005). Modeling of under-detection of cases in disease surveillance. *Annals of Epidemiology* **15**, 335–343.

Banerjee, T., Liu, Y. and Crews, D. C. (2016). Dietary patterns and ckd progression. *Blood Purification* **41**, 117–122.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with application in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1–20.

Bigogo, G., Audi, A., Aura, B., Aol, G., Breiman, R. F. and Feikin, D. R. (2010). Health-seeking patterns among participants of population-based morbidity surveillance in rural western Kenya: implications for calculating disease rates. *International Journal of Infectious Diseases* **14**, e967–973.

Caskey, F. J. (2016). Prevalence and incidence of renal disease in disadvantaged communities in Europe. *Clinical Nephrology* **86**, 34–36.

Caudill, S. B. and Mixon, F. G. (1995). Modeling household fertility decisions: estimation and testing of censored regression models for count data. *Empirical Economics* **20**, 183–196.

Chang, T. I., Lim, H., Park, C. H., Rhee, C. M., Kalantar-Zadeh, K., Kang, E. W., Kang, S. W. and Han, S. H. (2020, 2022/12/21). Association between income disparities and risk of chronic kidney disease: a nationwide cohort study of seven million adults in Korea. *Mayo Clinic Proceedings* **95**, 231–242.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* **104**, 1646–1660.

de Oliveira, G. L., Argiento, R., Loschi, R. H., Assunção, R. M., Ruggeri, F. and D'Elia Branco, M. (2022). Bias correction in clustered underreported data. *Bayesian Analysis* **17**, 95–126.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Lang, D. Temple and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* **26**, 403–413.

Dvorzak, M. and Wagner, H. (2016). Sparse Bayesian modeling of underreported count data. *Statistical Modelling* **16**, 24–46.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Fitzpatrick, A. L., Powe, N. R., Cooper, L. S., Ives, D. G. and Robbins, J. A. (2004). Barriers to health care access among the elderly and who perceives them. *American Journal of Public Health* **94**, 1788–1794.

Gibbons, C. L., Mangen, M. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M. *and others*. (2014). Measuring underreporting and underascertainment in infectious disease data sets: a comparison of methods. *BMC Public Health* **14**, 147.

Hart, J. T. (1971). The inverse care law. *The Lancet* **297**, 405–412.

Hossain, M. P., Goyder, E. C., Rigby, J. E. and El Nahas, M. (2009). CKD and poverty: a growing global challenge. *American Journal of Kidney Diseases* **53**, 166–174.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Johnson, N. L., Kotz, S. and Kemp, A. W. (2005). *Univariate Discrete Distributions*, Wiley Series in Probability and Statistics. Wiley.

Jug, M. (2014). Information revolution from data to policy action in low-income countries: how can innovation help? https://www.paris21.org/sites/default/files/PARIS21-Discussion Paper3-Innovation.pdf.

Krop, J. S., Coresh, J., Chambless, L. E., Shahar, E., Watson, R. L., Szklo, M. and Brancati, F. L. (1999). A community-based study of explanatory factors for the excess risk for early renal function decline in blacks vs whites with diabetes: the Atherosclerosis Risk in Communities study. *Archives of Internal Medicine* **159**, 1777–1783.

Li, T., Trivedi, P. K. and Guo, J. (2003). Modeling response bias in count: a structural approach with an application to the national crime victimization survey data. *Sociological Methods & Research* **31**, 514–544.

Lin, C. C., Bruinooge, S. S., Kirkwood, M. K., Olsen, C., Jemal, A., Bajorin, D., Giordano, S. H., Goldstein, M., Guadagnolo, B. A., Kosty, M., Hopkins, S., Yu, J. B., Arnone, A., Hanley, A., Stevens, S. *and others*. (2015). Association between geographic access to cancer care, insurance, and receipt of chemotherapy: geographic distribution of oncologists and travel distance. *Journal of Clinical Oncology* **33**, 3177–3185.

MacEachern, S. N. (1999). Dependent nonparametric processes. In: *ASA Proceedings of the Section on Bayesian Statistical Science*, Volume 1. Alexandria, VA: American Statistical Association, pp. 50–55.

Mallappallil, M., Friedman, E. A., Delano, B. G., McFarlane, S. I. and Salifu, M. O. (2014). Chronic kidney disease in the elderly: evaluation and management. *Clinical Practice (London, England)* **11**, 525–535.

Müller, P. and Quintana, F. A. (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference* **140**, 2801–2808.

Müller, P., Quintana, F. A., Jara, A. and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. New York: Springer.

Papadopoulos, G. and Santos Silva, J. M. C. (2012). Identification issues in some double-index models for non-negative data. *Economics Letters* **117**, 365–367.

Pesce, F., Pasculli, D., Pasculli, G., De Nicola, L., Cozzolino, M., Granata, A. and Gesualdo, L. (2022). "The Disease Awareness Innovation Network" for chronic kidney disease identification in general practice. *Journal of Nephrology* **35**, 2057–2065.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability* **25**.

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

Pontoriero, G., Pozzoni, P., Vecchio, L. D. and Locatelli, F. (2007). International Study of Health Care Organization and Financing for renal replacement therapy in Italy: an evolving reality. *International Journal of Health Care Finance and Economics* **7**(2-3), 201–215.

Quintana, F. A., Müller, P., Jara, A. and MacEachern, S. N. (2022). The Dependent Dirichlet Process and related models. *Statistical Science* **37**, 24–41.

Ren, L., Du, L., Carin, L. and Dunson, D. (2011). Logistic stick-breaking process. *Journal of Machine Learning Research* **12**, 203–239.

Rigon, T. and Durante, D. (2021). Tractable Bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference* **211**, 131–142.

Rodríguez, A., Dunson, D. B. and Gelfand, A. E. (2010). Latent stick-breaking processes. *Journal of the American Statistical Association* **105**, 647–659.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**, 145–177.

Shlipak, M. G., Tummalapalli, S. L., Boulware, L. E., Grams, M. E., Ix, J. H., Jha, V., Kengne, A., Madero, M., Mihaylova, B., Tangri, N. *and others*. (2021). The case for early identification and intervention of chronic kidney disease: conclusions from a kidney disease: Improving global outcomes (KDIGO) controversies conference. *Kidney International* **99**, 34–47.

Smart, N. A., Dieberg, G., Ladhani, M. and Titus, T. (2014). Early referral to specialist nephrology services for preventing the progression to end-stage kidney disease. *Cochrane Database of Systematic Reviews* **18**, 1–92.

Stoner, O., Economou, T. and da Silva, G. Drummond Marques. (2019). A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association* **114**, 1481–1492.

Toth, F. (2014). How health care regionalisation in Italy is widening the North-South gap. *Health Economics, Policy and Law* **9**, 231–249.

van Oostrom, S. H., Gijssen, R., Stirbu, I., Korevaar, J. C., Schellevis, F. G., Picavet, H. S. J. and Hoeymans, N. (2016). Time trends in prevalence of chronic diseases and multimorbidity not only due to aging: data from general practices and health surveys. *PLoS One* **11**, 1–14.

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.

Whittemore, A. S. and Gong, G. (1991). Poisson regression with misclassified counts: application to cervical cancer mortality rates. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **40**, 81–93.

Winkelmann, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics* **21**, 575–587.

Winkelmann, R. and Zimmermann, K. F. (1993). Poisson-logistic regression. *Working Paper No. 93–18*. Department of Economics, University of Munich.