Multilevel logistic cluster-weighted model for outcome evaluation in healthcare

Paolo Berta [*1] and Veronica Vinciotti[2]

[1]Department of Statistics and Quantitative Methods, University Bicocca-Milan (Italy)
[2]Department of Mathematics, Brunel University London (UK)

Abstract

In healthcare, multilevel models are typically used to evaluate hospitals' performance and to rank hospitals accordingly. While multilevel models capture the hierarchical structure in the data, such as the grouping of patients into hospitals, these models do not account for additional latent structures. In this paper, we develop a novel multilevel logistic cluster-weighted model which can predict a binary outcome, such as mortality within 30 days of discharge, while accounting both for known and latent structures of the data. We develop an Expectation-Maximization algorithm for parameter estimation and a parametric bootstrap approach for assessing the variability of the estimators. Using a rich dataset of the Lombardy (Italy) healthcare system and focussing on the two wards of cardiosurgery and medicine, we show how the proposed model detects, in both cases, two well-defined clusters within the patient to hospital hierarchical structure of the data. A comparison with standard multilevel and cluster-weighted approaches reveals a better fit of the proposed model and a greater insight into the structure of the data. We show how this can have implications in the resulting league tables and thus how the proposed model can be a useful tool for policy makers and healthcare managers to conduct hospital evaluations.

Keywords: Cluster-weighted models, model-based clustering, multilevel models, EM algorithm, healthcare, hospital evaluation.

1 Introduction

Statistical models are used by policy makers and healthcare managers in order to evaluate hospitals' performance. Multilevel models (also referred to as random-effects models or hierarchical models) are typically used for hospital performance evaluations: the data have a natural hierarchical structure, with patients nested into wards and hospitals, and there are often various other contextual variables, at individual and aggregate level, which are needed to predict helthcare outcomes of interest. The use of this approach in healthcare was pioneered by the seminal paper by Goldstein and Spiegelhalter [12] and subsequently used in many other studies, e.g. [9, 17, 18, 19, 26].

Multilevel models require the knowledge of the hierarchical structure in the data. In cases when this is unknown but one expects latent groups in the data, finite mixture models

[*]contact to: Department of Statistics and Quantitative Methods, University of Miano-Bicocca - Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy). Email: paolo.berta@unimib.it

allow to account for this heterogeneity in the outcome distribution by splitting the population into a finite number of relatively homogeneous classes [20]. Ingrassia et al. [14] have generalized this framework by introducing the so-called Cluster-Weighted Models (CWMs). Here, the joint density of the outcome and the covariates is clustered into groups. This results in a mixture of local models, which are represented by the conditional densities of the outcome given the covariates within a group. These densities are weighted both by the local densities of the covariates, which are typically not considered within standard mixture regression models, and by the usual mixing weights.

When both known and latent groups are to be modelled, a strand of research has proposed an extension of the mixture models to the multilevel setting in order to disentangle latent classes within the natural grouping in the data [3, 5, 11, 22, 28, 30]. Recently, Berta et al. [6] have extended CWMs to the multilevel framework in the case when the dependent variable is Gaussian. Since healthcare evaluations are typically based on binary outcomes, such as mortality, in this paper we propose a novel extension of CWM suited to binary outcomes. We develop the model in a rather general form where the covariates are of a mixed nature, namely continuous and categorical, and we propose a parametric bootstrap approach for building confidence intervals of the parameter estimates. Using a rich dataset from the Lombardy healthcare system, we show how the evaluation of the hospital performance is affected both by known and latent groupings in the data and how the final results are strongly impacted by the use of a model which accounts for this heterogeneity. The final hospital evaluation results are often presented in the form of league tables, where hospitals are sorted according to their quality. The use of an appropriate statistical model allows to accurately disentangle the significant differences between hospitals at this stage [12].

The paper is organized as follows: in Section 2 the new Multilevel Logistic Cluster-Weighted Model (ML-CWM) is introduced; in Section 3 inference of the proposed model is discussed including the generation of confidence intervals for the parameters; in Section 4 we show how this new approach can be used in a context of hospital evaluation; finally, in Section 5 we make some concluding remarks.

2 The Multilevel Logistic Cluster-Weighted Model

A cluster-weighted framework allows to estimate the joint probability of a random vector of covariates $\boldsymbol{X}$ and a binary dependent variable Y. Suppose that $\boldsymbol{X}$ and Y are defined in some finite space Ω with values in $R^d \times R$ and that Ω is partitioned into C clusters, say $\Omega_1, \ldots, \Omega_C$. Extending the CWMs to the multilevel framework allows to account for the fact that both the conditional distribution $Y|\boldsymbol{X}$ and the marginal distribution $\boldsymbol{X}$ depend on the C groups. In this way, the joint density of $(Y, \boldsymbol{X})$ can be described by a mixture of conditional densities $p(Y|\boldsymbol{X}, \Omega_c)$ weighted on the marginal densities $p(\boldsymbol{X}|\Omega_c)$ by the mixture's weights w_c.

Let Y be a binary variable, and let $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{V})$ be the set of covariates, where $\boldsymbol{U}$ is a vector of p continuous covariates and $\boldsymbol{V}$ is a vector of q categorical covariates, so that $d = q + p$. Assuming that $\boldsymbol{U}$ and $\boldsymbol{V}$ are locally independent within the clusters, a simplification usually adopted in model based clustering [13, 15, 29], the joint probability

can be factorized as

$$p\left(\boldsymbol{x}, y; \boldsymbol{\theta}\right) = \sum_{c=1}^{C} w_c p(y|\boldsymbol{x}; \boldsymbol{\xi}_c) p(\boldsymbol{x}; \boldsymbol{\nu}_c) =$$

$$= \sum_{c=1}^{C} w_c p(y|\boldsymbol{x}; \boldsymbol{\xi}_c) \phi(\boldsymbol{u}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \psi(\boldsymbol{v}; \boldsymbol{\lambda}_c), \tag{1}$$

where $\boldsymbol{\theta}$ is the vector of all model parameters, with $\boldsymbol{\xi}_c$ the vector of the cluster-dependent fixed and random effects of the regression part of the model and $\boldsymbol{\nu}_c$ the vector of cluster-dependent parameters for the covariate part of the model. In this paper, the density ϕ of $\boldsymbol{u}$ is assumed multivariate normal with mean $\boldsymbol{\mu}_c$ and covariance $\boldsymbol{\Sigma}_c$, whereas the density ψ of $\boldsymbol{v}$ is given by the product of q conditionally independent multinomial distributions with q parameters λ_{cr}, $r = 1, ..., q$, as espressed in [15]. For the covariance $\boldsymbol{\Sigma}_c$, we consider several parametrizations, as discussed in [25] and [21].

The conditional distribution of Y given $\boldsymbol{X}$ is Bernoulli, with a probability π which depends both on the hierarchical structure in the data (in our case patients within hospitals) and on possible latent groups. In particular, considering a logit link and a random effect model, the probability is described by

$$\text{logit}(\pi_j|X, C = c) = \alpha_c + \boldsymbol{\beta}_c \boldsymbol{X}_{cj} + b_{cj}, \tag{2}$$

where $b_{cj} \sim \mathcal{N}(0, \sigma^2_{bc})$ is the random effect for hospital j in the cluster c. For this case of a random effect model, the vector of parameters $\boldsymbol{\xi}_c$ in Equation 1 is given by $\boldsymbol{\xi}_c = (\alpha_c, \boldsymbol{\beta}_c, \sigma^2_{bc})$. Although a canonical logit link is used here, other links for binary outcomes can be considered at this stage.

This model extends the multilevel CWM proposed by Berta et al [6] to the case of a logistic link, which is particularly needed in the evaluation of healthcare systems where outcomes are typically binary. In the healthcare evaluation context, the random effects measure the relative effectiveness of hospitals with respect to the adjusted outcome under consideration. Taking mortality as the outcome, a positive and significant value of the random effect b_{cj} for the hospital j in the cluster c indicates that the hospital j has a negative effectiveness compared to the other evaluated hospitals within that cluster. At the opposite spectrum, a well performing hospital will be associated with a negative and significant random effect.

Using this model, which we denote by ML-CWM, each patient can be assigned to one of the C clusters according to the maximum posterior probability

$$p(\Omega_c|\boldsymbol{x}, y; \boldsymbol{\theta}) = \frac{w_c p(y|\boldsymbol{x}; \boldsymbol{\xi}_c) \phi(\boldsymbol{u}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \psi(\boldsymbol{v}; \boldsymbol{\lambda}_c)}{\sum_{k=1}^{C} w_k p(y|\boldsymbol{x}; \boldsymbol{\xi}_k) \phi(\boldsymbol{u}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \psi(\boldsymbol{v}; \boldsymbol{\lambda}_k)}.$$

For the case of Gaussian covariates only, conditions for identifiability of the model can be derived using the same approach of [15] for generalized cluster-weighted models.

3 Inference for ML-CWM

3.1 The EM-algorithm

In the presence of latent groups, the parameters $\boldsymbol{\theta}$ are estimated by an Expectation-Maximization (EM) algorithm, and the estimates of parameters which results from this method are those

that achieve at least a local maximum of the likelihood. Using the notation in Equation 1, the aim of the estimation process is to identify the vector of parameters $\boldsymbol{\theta}$ composed by

$$\boldsymbol{\theta} = (w_1 \dots w_{C-1}, \boldsymbol{\xi}_1 \dots \boldsymbol{\xi}_C, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_C, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_C, \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_C)',$$

The log-likelihood for $\boldsymbol{\theta}$ can be expressed as:

$$\ell((\boldsymbol{x}, y)|\boldsymbol{\theta}) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \log\{\sum_{c=1}^{C} p(y_{ij}|\boldsymbol{x}_{ij}; \boldsymbol{\xi}_c)\phi(\boldsymbol{u}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\psi(\boldsymbol{v}; \boldsymbol{\lambda}_c)\}$$

where y_{ij} and $\boldsymbol{x}_{ij}$ are the observed values of Y and $\boldsymbol{X}$, respectively, for the i^{th} first level observation (patient) in the j^{th} second level unit (hospital), with $j = 1, \dots, J$ and $i = 1, \dots, n_j$, where n_j is the total number of patients admitted to the hospital j.
Following [20], the formulation of the CWM problem can be viewed as a situation of incomplete data and an EM algorithm can be applied in order to estimate the maximum likelihood and to identify the probability that the observation $(\boldsymbol{x}_{ij}, y_{ij})$ belongs to one of the identified clusters. Assuming a C-dimensional component-label vector $\boldsymbol{z}_{ij}$ where $z_{ijc} = 1$ if the observation $(\boldsymbol{x}_{ij}, y_{ij})$ belongs to the c^{th} cluster and 0 otherwise, and considering the presence of continuous and categorical covariates, the complete data log-likelihood function for the observation $(\boldsymbol{x}_{ij}, y_{ij})$ and the latent allocation z_{ijc} can be expressed as:

$$\begin{aligned}
\ell_c((\boldsymbol{x}, y, \boldsymbol{z})|\boldsymbol{\theta}) = & \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{c=1}^{C} z_{ijc} \log(w_c) + \\
& \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{c=1}^{C} z_{ijc} \log[p(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\xi}_c)] + \\
& \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{c=1}^{C} z_{ijc} \log[\phi(\boldsymbol{u}_{ij}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)] + \\
& \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{c=1}^{C} z_{ijc} \log[\psi(\boldsymbol{v}_{ij}; \boldsymbol{\lambda}_c)].
\end{aligned} \tag{3}$$

Equation 3 implies that the complete data log-likelihood is composed of the four parts of the model: the probability of belonging to one of the clusters, the regression part of the outcome given the covariates, the marginal multivariate Gaussian distribution of the continuous covariates and the marginal distribution of the categorical covariates.

The EM algorithm follows an iterative process starting with an evaluation of the missing data based on the available data (E-step) and then a maximization of the expected log-likelihood (M-step). Assuming an unknown $\boldsymbol{z}$ vector, the $(r+1)^{\text{th}}$ iteration of the EM-algorithm is based on the expectation with respect to $\boldsymbol{z}$ of the complete data log-likelihood $\ell_c((\boldsymbol{x}, y, \boldsymbol{z})|\boldsymbol{\theta})$ in Equation 3, with $\boldsymbol{\theta}$ estimated at the r^{th} iteration, i.e.

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) = E_{\boldsymbol{z}|(\boldsymbol{x},y);\boldsymbol{\theta}^{(r)}}(\ell_c((\boldsymbol{x}, y, \boldsymbol{z})|\boldsymbol{\theta})).$$

This requires the calculation of the probability that the observation $(\boldsymbol{x}_{ij}, y_{ij})$ belongs to the c^{th} cluster, since

$$E(z_{ijc}|(\boldsymbol{x}, y), \boldsymbol{\theta}^{(r)}) = Pr\{z_{ijc} = 1|(\boldsymbol{x}, y), \boldsymbol{\theta}^{(r)}\} = \tau_c((\boldsymbol{x}, y), \boldsymbol{\theta}^{(r)}).$$

Given our proposed ML-CWM model,

$$\tau_c((\boldsymbol{x}_{ij}, y_{ij}), \boldsymbol{\theta}^{(r)}) = \frac{p(y_{ij}|\boldsymbol{x}_{ij}; \boldsymbol{\xi}_c^{(r)})\phi(\boldsymbol{u}_{ij}; \boldsymbol{\mu}_c^{(r)}, \boldsymbol{\Sigma}_c^{(r)})\phi(\boldsymbol{v}_{ij}; \boldsymbol{\lambda}_c^{(r)})w_c^{(r)}}{\sum_{k=1}^{C} p(y_{ij}|\boldsymbol{x}_{ij}; \boldsymbol{\xi}_k^{(r)})\phi(\boldsymbol{u}_{ij}; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}_k^{(r)})\phi(\boldsymbol{v}_{ij}; \boldsymbol{\lambda}_k^{(r)})w_k^{(r)}},$$

leading to

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) = &\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{c=1}^{C} \tau_c((\boldsymbol{x}_{ij}, y_{ij}), \boldsymbol{\theta}^{(r)})\{\log(w_c) + \log[p(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\xi}_c)] + \\ &+ \log[\phi(\boldsymbol{u}_{ij}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)] + \log[\phi(\boldsymbol{v}_{ij}; \boldsymbol{\lambda}_c)]\}. \end{aligned}$$

At this point the M-Step performs the estimation of the maximum likelihood, obtaining the new parameters $\boldsymbol{\theta}$ for the next iteration of the E-Step. The iterative process continues until a pre-defined convergence criterion is met. The convergence is guaranteed when the Aitken acceleration index [1] is lower than a defined threshold, which is typically set to 1e-04.

From a computational point of view, EM algorithms can be sensitive to the starting point. Several initialization strategies can be implemented (see, e.g., [7], [4]), and these are described in [15] for cluster-weighted models. As well as repeated random initializations, we will consider also using repeated k-means initializations, which make better use of the data available. The value maximizing the observed-data log-likelihood among these repeated initializations is selected.

The end of the EM algorithm provides two main results: the allocation of the observations to one of the identified clusters and the estimates of the parameters for both the fixed and the random effects of the regression part.

3.2 Standard errors via parametric bootstrap

Considering that the number of level 2 units could be small and that asymptotic MLE procedures may not accurately estimate the uncertainty associated with the parameter estimates from the EM algorithm [23], we include in the estimation process a further step consisting of a bootstrap process, in order to assess the variability of the EM estimates. In particular, we implement a parametric bootstrap approach for mixed models, adapting the steps described in [8] to the ML-CWM framework. Using the notation of Equation 2, and considering C latent clusters, the bootstrap approach follows these steps:

1. We denote by $\hat{\alpha}_c$, $\hat{\boldsymbol{\beta}}_c$, $\hat{\sigma}^2_{bc}$, $\hat{\boldsymbol{\mu}}_c$, $\hat{\boldsymbol{\Sigma}}_c$, $\hat{\boldsymbol{\lambda}}_c$, the set of the estimated parameters in the c-th latent cluster, and by $\hat{\tau}$ the group membership probabilities, from a model fitted on the full data using the EM algorithm described in the previous section.
2. We simulate the vector of the random effects $b_j^* \sim N(0, \hat{\sigma}^2_{bc})$, for $j = 1, \ldots, J$ and $c = 1, \ldots, C$;
3. We simulate the covariates $\boldsymbol{X}_c^*$ from a $N(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c)$ if continuous, and from a Multinomial with parameters $\hat{\boldsymbol{\lambda}}_c$ if categorical;
4. We simulate the bootstrap data y^*_{ijc} from a Bernoulli(π^*_{ijc}) with

$$\boldsymbol{\pi}^*_{cj} = \exp(\hat{\alpha_c} + \hat{\boldsymbol{\beta}}_c \boldsymbol{X}^*_{cj} + b^*_{cj})/(1 + \exp(\hat{\alpha_c} + \hat{\boldsymbol{\beta}}_c \boldsymbol{X}^*_{cj} + b^*_{cj}));$$

5. We refit the model on y^*_{ij}, using $\hat{\tau}$ as the initialization of the cluster allocations, and we obtain the set of bootstrap parameters $\hat{\alpha}^*_c$, $\hat{\boldsymbol{\beta}}^*_c$, $\hat{\sigma}^{2*}_{bc}$, $\hat{\boldsymbol{\mu}}^*_c$, $\hat{\boldsymbol{\Sigma}}^*_c$ and $\hat{\boldsymbol{\lambda}}^*_c$;

6. We repeat the steps 2-5 B times, where B is the number of bootstrap iterations.

Within the bootstrap iterative procedure, we initialize the EM algorithm using the posterior probabilities $\hat{\tau}$ estimated from the full data. This avoids possible problems with label switching and small sample sizes, as suggested by [23].

In order to evaluate whether this procedure produces the correct coverage error, we perform a simulation study on a dataset with a hierarchical structure and a different number of units at level 1 and level 2 which provide different degrees of imbalance. We consider a binary dependent variable and three covariates, two of which distributed as a standard normal and one categorical with two categories. We consider the case of 2 latent clusters and we fix the parameters as in Table 1, whereby the covariance matrix of the two continuous covariates is assumed diagonal and with equal variance (case "EII" in [25]). In order to measure the coverage of the bootstrap procedure, we construct 50 datasets from the same model and, in each case, we build bootstrap confidence intervals for each parameter using $B = 100$ bootstrap replications. For a single parameter, the coverage is defined by the percentage of times that the true parameter falls in the confidence intervals. If the confidence level is set to 95%, one would expect the coverage close to this nominal value [8]. Table 2 shows a good performance in terms of coverage for the bootstrap algorithm and shows that the higher is the number of observations the better the coverage.

	Cluster 1	Cluster 2
α	-0.5	0.5
β_1	0.5	-0.5
β_2	0.1	-0.1
β_3	0.5	-0.5
λ	0.7	0.3
σ_b^2	4	2
μ_1	-4	-5
μ_2	4	5
$(\Sigma)_{i,i}$	1	1

Table 1: Parameters used for data generation from the ML-CWM model.

4 Hospital evaluation by a ML-CWM approach

We analyze an administrative dataset gathered from the Lombardy region in Italy, which collects information on patients admitted to 150 hospitals in 2014. The data include demographic information (age, gender), information on hospitalizations, such as length of stay, special-care unit use, within-hospital mortality, and a total of 6 diagnosis codes and procedures defined according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). In order to test the ability of the model in identifying clusters among patients in the context of prediction of hospital performance within a specific discipline, we test the ML-CWM on two different disciplines: cardiosurgery and medicine. Cardiosurgery is a highly specialized discipline admitting patients that need complex surgical intervention, but with a low risk of 30-day mortality, whereas medicine is a widespread general discipline, characterized by a high risk of mortality.

Parameters	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
	20 level 1, 15 level 2		40 level 1, 20 level 2		100 level 1, 30 level 2	
α	0.92	0.90	0.90	0.92	0.93	0.94
β_1	0.96	0.90	0.92	0.94	0.97	0.94
β_2	0.96	0.98	0.98	0.94	0.94	0.93
β_3	0.96	0.94	0.94	0.94	0.94	0.92
σ_b^2	0.86	0.86	0.86	0.84	0.88	0.89
μ_1	0.90	0.91	0.92	0.92	0.91	0.92
μ_2	0.94	0.90	0.91	0.92	0.92	0.92
Σ	0.92	0.89	0.94	0.98	0.92	0.94
λ	0.96	0.92	0.92	0.98	0.93	0.94
Observations w/i clusters	300	300	800	800	3000	3000
Num. of Observations	600		1600		6000	

Table 2: Coverage of the parametric bootstrap approach for 95% confidence intervals from simulated data with varying sample sizes.

The outcome of interest is 30-day mortality, the most used proxy of quality in this research field. This outcome is measured by merging the hospital record described above with the registry of citizens conserved in Lombardy, where we can find the date of death for each patient. In this way we can identify whether a patient discharged alive has died within 30 days after the discharge. In the event of death for patient i, the outcome is recorded as $y_i = 1$.

A number of characteristics are selected for each patient, namely sex, age, the DRG weight, measuring the resources used by the hospital to treat each patient, and the Elixhauser index [10], measuring the level of patients' comorbidities. These patients' characteristics are typically used in healthcare evaluation frameworks as risk-adjustment covariates, and age and gender in particular are notorious for accounting for the largest share of case-mix variability while not being dependent on the quality provided during the hospitalization [2, 24].

We apply the proposed ML-CWM to this dataset, separately for cardiosurgery hospitalizations and for patients admitted in medicine. We exploit the hierarchical structure of patients nested within hospitals using a multilevel model and we use the proposed multilevel cluster-weighted model to investigate whether there is evidence for further latent structures. To initialise the clusters, we use repeated k-means initializations: we found the cluster-weighted model to be more stable and faster under this initialization strategy than using random initializations. In particular, we performed 10 k-means initializations which resulted in 3 different starting points, but led to the same final estimates from the cluster-weighted model. Table 3 reports the results from the best fitting model for the case of two clusters. We fitted models under a growing number of latent clusters, but the inference failed with $C > 2$ even after various attempts of tweaking parameters and options. According to the descriptive statistics in Table 3, the covariate age disentangles very well the latent hetereogeneity in the data and when 2 clusters are identified (young patients and older patients), there are no other sensible divisions. Both for the case of one cluster ($C = 1$, standard multilevel model) and for the case of two clusters ($C = 2$), we tested a large number of parametrizations of the covariance matrix $\mathbf{\Sigma}_c$ for the three continuous covariates. Using the nomenclature of [21], we considered the "EEI", "VII", "EEI", "VEI", "EVI", "VVI", "EEE", "VVV", "EVV", "VEV", "EVE", "VEE", "VVE", and "VVV" structures

for $C = 2$ and the "EII", "EEI", and "EEE" structures for $C = 1$. For both cardiosurgery and medicine the case of two clusters and the "VEV" decomposition performed best, by giving the lowest Bayesian Information Criterion (BIC) [27], namely 145,055.8 for cardiosurgery and 1,981,374 for medicine.

The descriptive statistics in Table 3 from the fitted model allow us to appreciate the different case-mix of patients admitted in the two considered wards and in the two identified clusters. Patients in cardiosurgery are on average younger than patients in medicine, while the risk of mortality in medicine is 10 times higher compared to cardiosurgery. The value of the DRG weight shows how cardiosurgery is a highly specialized discipline. The clustering composition for cardiosurgery indicates how the two latent groups mainly differ according to the age (in cluster 2 the patients are younger). Considering the clusters identified in medicine, we observe that patients in cluster 2 are younger than patients in cluster 1, which leads to a lower risk of mortality for this cluster compared to cluster 1. Moreover, higher levels of comorbidities are observed for the patients allocated to cluster 1. The last column compares the observed mortality in the identified clusters with the expected mortality by the fitted ML-CWM model. The expected rates are quite similar to the observed ones, indicating a good fit of the models and a need to include the hierarchical structure of the data in the model specification.

		Covariates				Outcome	
		DRG Weight	Comorbidities	Age	Female	Observed Mortality	Expected Mortality
Cardiosurgery							
Cluster 1	Mean	5.4110	1.2235	70.5777	0.3305	0.0140	0.0131
(#Obs 7,788)	Std Dev	2.8120	1.0693	8.4312	0.4704		
(#Hospitals 20)							
Cluster 2	Mean	5.5688	1.0551	42.6457	0.3363	0.0067	0.0053
(#Obs 889)	Std Dev	3.3237	0.8868	8.4884	0.4727		
(#Hospitals 20)							
Medicine							
Cluster 1	Mean	1.0922	1.3463	79.0967	0.5179	0.1660	0.1641
(#Obs 119,678)	Std Dev	0.7092	1.1230	9.1036	0.4997		
(#Hospitals 107)							
Cluster 2	Mean	0.9399	0.8611	44.0957	0.4851	0.0605	0.0526
(#Obs 18,407)	Std Dev	0.5520	0.9337	10.3680	0.4998		
(#Hospitals 107)							

Table 3: Descriptive statistics of the two clusters identified by ML-CWM on the two separate wards of cardiosurgery and medicine.

We compare the results provided by the ML-CWM with a classical multilevel model in terms of goodness of fit and parameter estimates of both the fixed and random effects. In addition, in order to check whether the main improvement of the ML-CWM comes from the multilevel or from the cluster-weighted aspect, we also compare the results with a standard logistic CWM (L-CWM), which does not contain the random effects in the model and thus does not consider the hierarchical structure in the data. In each case, we selected

the optimal parametrization of the covariance of the continuous covariates based on BIC. This resulted in the choice of the "EEE" decomposition both for the multilevel and the L-CWM models. Table 4 shows that the ML-CWM has the lowest BIC, compared to both the classical multilevel model and the standard logistic CWM. Looking further at the parameter estimates, Table 4 shows how several effects of the covariates on the risk of mortality are different between latent groups, and how these differences are not picked up by the standard multilevel model. The significance level of the estimates is evaluated via the parametric bootstrap approach described before. For cardiosurgery, the model finds a different direction and significance for the effect of age and DRG weight on mortality among the two clusters. Whereas in the standard multilevel model the coefficient related to the DRG weight is negative and significant, the application of both the CWM and ML-CWM shows how this relationship is positive and significant for the patients allocated in cluster 1, indicating that the higher the resources used by the hospital to treat patients in this group the higher their risk of mortality, while the relationship is negative and not-significant for the patients allocated in cluster 2. Moreover, we observe that cluster 2 is characterized by non-significant coefficients for the patients' gender and that the estimated coefficient for comorbidities is positive and significant only in cluster 1. The covariate related to the patients' age is, as expected, always positive and significant, confirming the main role of this covariate in adjusting the risk of mortality in our analysis. In particular, this covariate is the only covariate explaining the differences in the risk of mortality for the patients allocated in cluster 2. In contrast to this, in the ward of medicine, we do not observe any differences in the direction of coefficients in the compared models, but using the ML-CWM we detect a different magnitude for the coefficients related to age and DRG weight. In particular in cluster 1 the DRG weight has the highest magnitude compared to the other coefficients. Finally, in medicine we observe a gender effect in cluster 1, with female patients having a lower risk of mortality in that cluster. The results show the flexibility of the model in capturing the impact of variables on the adjusted outcome when latent structures are present: covariates such as gender in cardiosurgery do not have an impact on mortality within any of the identified clusters, corresponding to the case of a variable that could be dropped from the model. Other covariates on the other hand, have an impact on the outcome only for some of the clusters or a different impact across the clusters, a situation that could not be contemplated by a traditional multilevel model.

The effects detected in Table 4 have a significant impact on the final league tables, and show also here a difference between the results obtained by the proposed ML-CWM model and by a standard multilevel approach. Figure 1 shows the league tables for cardiosurgery using the multilevel model, in the first top plot, and the ML-CWM at the bottom. Figure 2 provides the same results for medicine. These figures are drawn based on the estimated random effects and on confidence intervals obtained using the same parametric bootstrap approach described before. Such plots can be produced, for each cluster, using the function `plotREsim` in the R package `merTools`, where we use the option of plotting in the odds ratio scale [16]. Hospital random effects different from the overall average (i.e. when the confidence interval does not cross the red line) are highlighted in bold. The figures show how, in cardiosurgery, the league tables of the multilevel model and of cluster 1 of ML-CWM are the same, but ML-CWM allows to detect a bad performance related to the hospital coded as `7` in cluster 2 . This is the only hospital presenting a bad performance in this cluster, and it is the only hospital with bad results both in cluster 1 and cluster 2. This means that the patients allocated in cluster 2 receive the same quality in all the hospitals except for hospital `7` (see last plot in Figure 1).
In medicine, we are able to compare the overall heterogeneity of the first plot in Figure 2

	Covariates	Multilevel	L-CWM		ML-CWM	
Cardiosurgery			Cluster1	Cluster2	Cluster1	Cluster2
	(Intercept)	-8.5821***	-8.5868***	-8.5097***	-8.5509***	-30.0003***
	Female	0.1786	0.2142	0.3393	0.2045	9.1940
	Age	0.0517***	0.0676***	0.0467***	0.1319***	0.2299**
	DRG Weight	-0.0339***	0.1353**	-0.098***	0.437**	-0.8226***
	Elix Index	0.2485***	0.6892***	0.197	0.1318***	-10.4423
	#Parameters	6	37		39	
	BIC	147,644.7	145,255.8		145,055.8	
Medicine						
	(Intercept)	-5.7630***	-6.7061***	-6.7950***	-6.8485***	-7.1120***
	Female	-0.2683***	-0.2559***	0.0097	-0.2515***	-0.0208
	Age	0.0491***	0.0442***	0.0626***	0.0436***	0.0678***
	DRG Weight	0.3679***	1.6026***	0.1365***	1.6780***	0.1255***
	Elix Index	0.0677***	-0.0099	-0.0682	0.0212***	-0.0579
	#Parameters	6	37		39	
	BIC	2,149,655	1,982,832		1,981,374	

Table 4: Regression coefficients of the multilevel, the L-CWM and the proposed ML-CWM models fitted to the Lombardy healthcare data for the cases of cardiosurgery and medicine. */**/*** indicate 10%/5%/1% significance from the the boostrap procedure.

with the cluster specific heterogeneity in the last two plots in Figure 2. As we observed for cardiosurgery, patients allocated in cluster 2 receive a more homogeneous level of quality. This could therefore be a useful cluster for policy makers to identify the hospitals that have a bad performance for a specific analysis. Furthermore, we can detect a number of hospitals, i.e. "8", "29", "90", "40", which are ranked as having a significant bad performance in both clusters. However, there are several other hospitals, such as "26", "54", "46", "83" and others, which are bad performers in the first cluster but they are ranked as average in the second one, showing a greater flexibility in ranking when latent clusters are accounted for. The final ranking of the hospital can be obtained using the expected rank test, which is implemented in the `expectedRank` function in the R package `merTools`. The ranking produced by this test confirmed the best and worst hospitals identified above.

5 Conclusions

In this paper we have presented an extension of multilevel cluster-weighted models for binary outcomes and have shown its use and benefits within a hospital evaluation framework. The proposed model allows to identify latent clusters in the data, related to both the outcome and the risk-adjustment variables, as well as to account for the hierarchical structure of the data which is typical in healthcare evaluations. We present inference of the model, including an EM-algorithm for parameter estimation and a parametric bootstrap approach for building confidence intervals for the parameters.

Using a rich dataset on the Lombardy healthcare system, we show how the proposed multilevel cluster-weighted model detects two well-defined latent groups within the hierarchical structure of hospitals. Interestingly, the regression coefficients have different signs,

Figure 1: League Tables for the Multilevel Model (first) and ML-CWM (second and third) in Cardiosurgery

Figure 2: League Tables for the Multilevel Model (first) and ML-CWM (second and third) in Medicine

magnitude and statistical significance for the two different groups, showing the advantage of this method compared to a standard multilevel model. The Bayesian information criterion supports this comparison. In addition to the fixed effects, the league tables of hospitals constructed from the random effects show different patterns between the two latent groups. This may have great implications for policy makers and healthcare managers because these effects could be masked using a classic approach and the final rankings of hospitals may be biased.

This paper provides a new method to evaluate performance in the healthcare sector. However, the proposed model can be widely applied in all research fields where there is a binary outcome and a hierarchical structure of the data. For example, education is a typical field of research where data are characterized by a hierarchical structure and binary outcomes are often considered.

References

[1] A Aitken, *On Bernoulli's numerical solution of algebraic equations*, Proceedings of the Royal Society of Edinburgh **46** (1926), 289–305.

[2] Arlene S Ash, Stephen F Fienberg, Thomas A Louis, Sharon-Lise T Normand, Therese A Stukel, and Jessica Utts, *Statistical issues in assessing hospital performance*, (2012).

[3] T Asparouhov and B Muthén, *Advances in latent variable mixture models*, Advances in Latent Variable Mixture Models (Charlotte) (G.R. Hancock and K.M. Samuelson, eds.), Information Age Publishing, 2008, pp. 27–51.

[4] Luca Bagnato and Antonio Punzo, *Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm*, Computational Statistics **28** (2013), no. 4, 1571–1597.

[5] F Bartolucci, F Pennoni, and G Vittadini, *Assessment of school performance through a multilevel latent Markov Rasch model*, Journal of Educational and Behavioral Statistics **36** (2011), no. 4, 491–522.

[6] P Berta, S Ingrassia, A Punzo, and G Vittadini, *Multilevel cluster-weighted models for the evaluation of hospitals*, Metron **74** (2016), 275–292.

[7] Christophe Biernacki, Gilles Celeux, and Gérard Govaert, *Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models*, Computational Statistics & Data Analysis **41** (2003), no. 3-4, 561–575.

[8] J Carpenter, H Goldstein, and J Rasbash, *A novel bootstrap procedure for assessing the relationship between class size and achievement*, Journal of the Royal Statistical Society: Series C (Applied Statistics) **52** (2003), no. 4, 431–443.

[9] C Christiansen and C Morris, *Improving the statistical approach to health care provider profiling*, Annals of internal medicine **127** (1997), no. 8(2), 764–768.

[10] A Elixhauser, C Steiner, D Harris, and R Coffey, *Comorbidity measures for use with administrative data*, Medical care **36** (1998), no. 1, 8–27.

[11] M Gnaldi, S Bacci, and F Bartolucci, *A multilevel finite mixture item response model to cluster examinees and schools*, Advances in Data Analysis and Classification **10** (2016), no. 1, 53–70.

[12] H Goldstein and D Spiegelhalter, *League tables and their limitations: statistical issues in comparisons of institutional performance*, Journal of the Royal Statistical Society. Series A (Statistics in Society) (1996), 385–443.

[13] Christian Hennig and Tim F Liao, *How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification*, Journal of the Royal Statistical Society: Series C (Applied Statistics) **62** (2013), no. 3, 309–369.

[14] S Ingrassia, S Minotti, and G Vittadini, *Local statistical modeling via a cluster-weighted approach with elliptical distributions*, Journal of classification **29** (2012), no. 3, 363–401.

[15] S. Ingrassia, A Punzo, G. Vittadini, and S. C. Minotti, *The generalized linear mixed cluster-weighted model*, Journal of Classification **32** (2015), no. 1, 85–113.

[16] J Knowles and C Frederick, *mertools: Tools for analyzing mixed effect regression models*, 2016, R package version 0.3.0.

[17] AH Leyland and FA Boddy, *League tables and acute myocardial infarction*, The Lancet **351** (1998), no. 9102, 555–558.

[18] AH Leyland and H Goldstein, *Multilevel modelling of health statistics*, Wiley, 2001.

[19] E Marshall and D Spiegelhalter, *Institutional performance*, Multilevel modelling of health statistics (2001), 127–142.

[20] G McLachlan and D Peel, *Finite mixture models*, John Wiley & Sons, New York, 2000.

[21] K Murphy and TB Murphy, *Gaussian parsimonious clustering models with covariates*, arXiv preprint arXiv:1711.05632v2 (2018).

[22] B Muthén and T Asparouhov, *Multilevel regression mixture analysis*, Journal of the Royal Statistical Society: Series A (Statistics in Society) **172** (2009), no. 3, 639–657.

[23] A O'Hagan, TB Murphy, L Scrucca, and IC Gormley, *Investigation of parameter uncertainty in clustering using a gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap*, arXiv preprint arXiv:1510.00551v4 (2018).

[24] Carol Propper and John Van Reenen, *Can pay regulation kill? panel data evidence on the effect of labor markets on hospital performance*, Journal of Political Economy **118** (2010), no. 2, 222–273.

[25] A E Raftery and N Dean, *Variable selection for model-based clustering*, Journal of the American Statistical Association **101** (2006), no. 473, 168–178.

[26] N Rice and A Leyland, *Multilevel models: applications to health data*, Journal of Health Services Research **1** (1996), no. 3, 154–164.

[27] G Schwarz, *Estimating the dimension of a model*, The annals of statistics **6** (1978), no. 2, 461–464.

[28] J Vermunt, *Multilevel latent class models*, Sociological methodology **33** (2003), no. 1, 213–239.

[29] Jeroen K Vermunt and Jay Magidson, *Latent class cluster analysis*, vol. 11, Cambridge University Press, 2002.

[30] JK Vermunt, *Mixed-effects logistic regression models for indirectly observed discrete outcome variables*, Multivariate Behavioral Research **40** (2005), no. 3, 281–301.