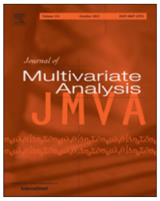
Journal of Multivariate Analysis 168 (2018) 63–74

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva



Classified mixed logistic model prediction

Hanmei Sun [a], Thuan Nguyen [b], Yihui Luan [a], Jiming Jiang [c,*]

[a] School of Mathematics, Shandong University, Jinan, Shandong 250100, China
[b] Oregon Health and Science University, Portland, OR 97239-3098, USA
[c] Department of Statistics, University of California, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 4 February 2018
Available online 30 June 2018

AMS subject classifications:
62J12
62M20
62F40

Keywords:
Clustered binary data
CMLMP
CMMP
Matching
Mixed logistic model
Mixed model prediction
MSPE

ABSTRACT

We develop a classified mixed logistic model prediction (CMLMP) method for clustered binary data by extending a method proposed by Jiang et al. (2018) for continuous outcome data. By identifying a class, or cluster, that the new observations belong to, we are able to improve the prediction accuracy of a probabilistic mixed effect associated with a future observation over the traditional method of logistic regression and mixed model prediction without matching the class. Furthermore, we develop a new strategy for identifying the class for the new observations by utilizing covariates information, which improves accuracy of the class identification. In addition, we develop a method of obtaining second-order unbiased estimators of the mean squared prediction errors (MSPEs) for CMLMP, which are used to provide measures of uncertainty. We prove consistency of CMLMP, and demonstrate finite-sample performance of CMLMP via simulation studies. Our results show that the proposed CMLMP method outperforms the traditional methods in terms of predictive performance. An application to medical data is discussed.

1. Introduction

Many practical problems are related to prediction of characteristics of interest at subject (e.g., precision medicine) or subpopulation (e.g., precision public health) level. In such cases, it is possible to make substantial gains in prediction accuracy by identifying a class that a new subject belongs to. This was recently demonstrated by Jiang et al. [6], who proposed a method called classified mixed model prediction (CMMP). The idea is to create a "match" between the classes, or clusters, of the training data and the potential class of the new data. Once such a class is identified, a mixed model prediction (MMP) technique can be used to improve prediction accuracy; see, e.g., Section 2.3 in [3].

The CMMP method developed in [6] applies only to linear models for continuous responses. However, clustered binary data frequently occur in practice. For example, Thromboembolic or hemorrhagic complications [2] occur in as many as 60% of patients who underwent extracorporeal membrane oxygenation (ECMO), an invasive technology used to support children during periods of reversible heart or lung failure [7]. Over half of pediatric patients on ECMO are currently receiving antithrombin (AT) to maximize heparin sensitivity. In a retrospective, multi-center, cohort study of children (≤ 18 years of age) who underwent ECMO between 2003 and 2012, 8601 subjects participated in 42 free-standing children's hospitals across 27 US states and the District of Columbia known as Pediatric Health Information System (PHIS). Data were de-identified prior to inclusion in the study dataset; however, encrypted medical record numbers allowed for tracking of individuals across multiple hospitalizations. Many of the outcome variables were binary, such as the bleed_binary variable, which is a main outcome variable indicating hemorrhage complication of the treatment; and the DischargeMortalit1Flag

---

* Corresponding author.
E-mail address: jiang@wald.ucdavis.edu (J. Jiang).

https://doi.org/10.1016/j.jmva.2018.06.004

variable, which is associated with mortality. Here the treatment refers to AT. Prediction of characteristics of interest associated with the binary outcomes, such as probabilities of hemorrhage complication or those of mortality for specific patients are of considerable interest. Note that the data are also potentially clustered, with the clusters corresponding to the children's hospitals. In addition to the treatment indicator, there were 20 other covariate variables, for which information were available. More details about the data will be provided in Section 5.

For clustered binary data, such as in the ECMO example, Jiang and Lahiri [4] proposed an empirical best prediction approach for predicting a mixed effect, such as the probability of hemorrhage complication in the ECMO example. Also see Section 3.6.2 in [3]. The method is based on a mixed logistic model, which assumes that, given the subject-specific random effects, $\alpha_1, \ldots, \alpha_m$, binary responses y_{ij} with $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n_i\}$ are conditionally independent with the conditional probability satisfying

$$\Pr(y_{ij} = 1|\alpha) = p_{ij} \quad \text{with} \quad \text{logit}(p_{ij}) = x_{ij}^\top \beta + \alpha_i, \tag{1}$$

where $\text{logit}(p) = \ln\{p/(1-p)\}$. Here i is the index for subject (e.g., patient, or group of patients), j is the index for observation within the subject (e.g., observation collected at the jth time point, or observation collected from the j patient in the subject group); x_{ij} is a vector of observed covariates, and β is a vector of unknown fixed effects. Furthermore, $\alpha_1, \ldots, \alpha_m$ are random effects, assumed to be independent and distributed as $\mathcal{N}(0, \sigma^2)$, where σ^2 is an unknown variance. For any known vector x and function g, the best predictor (BP) of the mixed effect, $\theta = g(x^\top \beta + \alpha_i)$, in the sense of minimum mean squared prediction error (MSPE), is given by

$$\tilde{\theta} = \text{E}(\theta|y) = \frac{\text{E}[g(x^\top \beta + \sigma\xi)\exp\{y_{i\cdot}\sigma\xi - \sum_{j=1}^{n_i} \ln(1 + e^{x_{ij}^\top \beta + \sigma\xi})\}]}{\text{E}[\exp\{y_{i\cdot}\sigma\xi - \sum_{j=1}^{n_i} \ln(1 + e^{x_{ij}^\top \beta + \sigma\xi})\}]}, \tag{2}$$

where $y_{i\cdot} = y_{i1} + \cdots + y_{in_i}$, and the expectations are taken with respect to $\xi \sim \mathcal{N}(0, 1)$. Two special cases of Eq. (2) are the following:

(a) If the covariates are at the cluster level, i.e., $x_{ij} = x_i$, and $g(u) = \text{logit}^{-1}(u) = e^u/(1 + e^u)$, then, (2) reduces to

$$\tilde{p} = \frac{\text{E}[\text{logit}^{-1}(x^\top \beta + \sigma\xi)\exp\{y_{i\cdot}\sigma\xi - n_i \ln(1 + e^{x_i^\top \beta + \sigma\xi})\}]}{\text{E}[\exp\{y_{i\cdot}\sigma\xi - n_i \ln(1 + e^{x_i^\top \beta + \sigma\xi})\}]}, \tag{3}$$

which is the BP of $p = \text{logit}^{-1}(x^\top \beta + \alpha_i)$. Note that, in this case, the mixed effect is a subject-specific (conditional) probability, such as the probability of hemorrhage complication of the AT treatment in the ECMO problem discussed above, for a specific patient.

(b) If $x = 0$, and $g(u) = u$, Eq. (2) reduces to

$$\tilde{\alpha}_i = \sigma \frac{\text{E}[\xi \exp\{y_{i\cdot}\sigma\xi - \sum_{j=1}^{n_i} \ln(1 + e^{x_{ij}^\top \beta + \sigma\xi})\}]}{\text{E}[\exp\{y_{i\cdot}\sigma\xi - \sum_{j=1}^{n_i} \ln(1 + e^{x_{ij}^\top \beta + \sigma\xi})\}]}, \tag{4}$$

which is the BP of α_i, the subject-specific (e.g., hospital) random effect. The special cases (3) and (4) will be used in the sequel.

In Eqs. (2), (3), or (4), β and σ are understood as the true parameters, which are typically unknown in practice. It is then customary to replace β, σ by their consistent estimators. The results are called empirical BP, or EBP. In this paper, we assume that the sample size for the "training data" is sufficiently large that the EBP is approximately equal to the BP [4]. Hereafter, by training data we refer to the data $\{y_{ij} : 1 \leq i \leq m, 1 \leq j \leq n_i\}$ described above that satisfy the assumed mixed logistic model.

Our main interest is to predict a mixed effect that is associated with a set of new observations. More specifically, let the new, binary observations be $y_{n,1}, \ldots, y_{n,n_{\text{new}}}$, and the corresponding covariates be $x_{n,1}, \ldots, x_{n,n_{\text{new}}}$ such that, conditional on a random effect α_I that has the same $\mathcal{N}(0, \sigma^2)$ distribution, $y_{n,1}, \ldots, y_{n,n_{\text{new}}}$ are mutually independent with

$$\Pr(y_{n,k} = 1|\alpha_I) = p_{n,k} \quad \text{and} \quad \text{logit}(p_{n,k}) = x_{n,k}^\top \beta + \alpha_I, \tag{5}$$

where β is the same as in Eq. (1). Typically, the sample size, n_{new}, for the new observations is limited. If one relies only on the new observations to estimate the mixed effect, say, $p_{n,k}$ for a given k, the available information is limited. Luckily, one has much more than just the new observations. It would be beneficial if one could "borrow strength" from the training data, which are much larger in size. For example, if one knows that $I = i$, then, there is a much larger cluster in the training data, namely, $y_{i1}, \ldots, y_{in_i}$, corresponding to the same cluster-specific random effect, α_I. This cluster in the training data is much larger because, quite often, n_i is much larger than n_{new}. One can also utilize the training data to estimate the unknown parameters, β and σ, which would be much more accurate than using only the new observations. As noted, with accurate estimation of the parameters, the EBP will closely approximate the BP [4]. Thus, potentially one has a lot more information that can be used to estimate the mixed effect of interest associated with α_I. The difficulty is, however, that I is unknown. In fact, at this point, one does not know the answer to any of the following questions: (a) Is there a "match" between I and one $i \in \{1, \ldots, m\}$ corresponding to the training data clusters? and (b) If there is, which one?

In Section 2, we develop a method that gives answer to question (b). As in [6], it turns out that the answer to question (a) does not matter much, so far as prediction of the mixed effect is concerned. In other words, even if the actual match does not exist, a CMMP procedure based on the false match still helps in improving prediction accuracy of the mixed effects. Once the matching class is identified, the EBP can be utilized, as noted above, to make more accurate prediction about the mixed effect of interest. We call this procedure classified mixed logistic model prediction (CMLMP) by an obvious analogy with CMMP. Consistency of CMLMP is established under both the matched and unmatched scenarios, and demonstrated empirically via simulation studies. Furthermore, in Section 3, we propose a new matching strategy that utilizes cluster-level covariates, when such information is available, and show that it improves accuracy of the mixed effect prediction. This is something not considered by [6], and a new contribution of the current paper. In particular, our simulation studies show that the new CMLMP strategy outperforms the standard logistic regression prediction (SLRP) as well as mixed logistic model prediction (MLMP) without matching the class, in which case the class of the new observations is assumed to be independent with the classes of the training data. In Section 4 we develop measures of uncertainty by proposing a second-order unbiased estimator of the MSPE of CMLMP. In Section 5, we revisit the ECMO data mentioned above and carry out a real-data analysis to further demonstrate empirical performance of CMLMP. Some concluding remarks are offered in Section 6. Proofs of the theoretical results are deferred to an Online Supplement.

2. Classified mixed logistic model prediction

In this section, we develop an extension of CMMP to binary data under the mixed logistic model described in the previous section, known as CMLMP. We begin by first considering a special case, in which the covariates are at the cluster level. Extension to the general case, in which the covariates are at the unit level, will be considered next.

2.1. A special case: Cluster-level covariates

By cluster-level covariates, it means that $x_{ij} = x_i$ for all i, j. Similarly, the covariates for the new observations are also at the cluster level, i.e., $x_{n,k} = x_n$.

2.1.1. The matched case

First assume that there is a match between I, the index for the random effect associated with the new observations, and one of the indexes, $i \in \{1, \ldots, m\}$, associated with the training-data random effects. However, this match is unknown to us. Thus, as a first step, we need to identify the match, i.e., an index $\hat{I} \in \{1, \ldots, m\}$ computed from the data, which may be viewed as an estimator of I.

Suppose that $I = i$. Then, by Eq. (3), the BP of $p_n = \Pr(y_{n,k} = 1|\alpha_I) = \mathrm{logit}^{-1}(x_n^\top \beta + \alpha_I) = \mathrm{logit}^{-1}(x_n^\top \beta + \alpha_i)$ is

$$\tilde{p}_{n,i} = \frac{\mathrm{E}[\mathrm{logit}^{-1}(x_n^\top \beta + \sigma\xi)\exp\{y_{i\cdot}\sigma\xi - n_i \ln(1 + e^{x_i^\top \beta + \sigma\xi})\}]}{\mathrm{E}[\exp\{y_{i\cdot}\sigma\xi - n_i \ln(1 + e^{x_i^\top \beta + \sigma\xi})\}]}. \tag{6}$$

In (6), the parameters β, σ are understood as the true parameters, which are unknown in practice. If we replace these parameters by their consistent estimators, such as the maximum likelihood (ML) or GEE estimators (see, e.g., Section 4.2 in [3]) based on the training data, we obtain the EBP of p_n, denoted by $\hat{p}_{n,i}$. Furthermore, an "observed" p_n is the sample proportion, $\bar{y}_n = (y_{n,1} + \cdots + y_{n,n_{\mathrm{new}}})/n_{\mathrm{new}}$. Our idea is to identify I as the index $i \in \{1, \ldots, m\}$ that minimizes the distance between $\hat{p}_{n,i}$ and $\bar{y}_n$, i.e.,

$$\hat{I} = \underset{i \in \{1,\ldots,m\}}{\mathrm{argmin}} |\hat{p}_{n,i} - \bar{y}_n|. \tag{7}$$

Given $\hat{I}$, the classified mixed logistic model predictor (CMLMP) of p_n is $\hat{p}_{n,\hat{I}}$.

The following theorem establishes consistency of CMLMP under suitable conditions. The proof is given in the Online Supplement.

Theorem 1. *Suppose that, in addition to the assumptions about the training data and the new observations, $|x_1|, \ldots, |x_m|$ are bounded. If $\hat{\beta}, \hat{\sigma}$ are consistent estimators of β, σ, respectively, then $\hat{p}_{n,\hat{I}} - p_n \xrightarrow{\Pr} 0$ as m, $\min(n_1, \ldots, n_m)$, and $n_{\mathrm{new}} \to \infty$. In other words, the CMLMP of p_n is consistent.*

2.1.2. Matched-or-unmatched case

In practice, one may not be sure whether there is a match between the random effect corresponding to the new observations and one of the random effects associated with the training data. This is what we call matched-or-unmatched

(M/UM) case. When there is no match, the index I is understood as an integer larger than m. This implies, in particular, that α_I and the new observations are independent with $\alpha_1, \ldots, \alpha_m$ and the training data. Following [6], we can extend the CMLMP method developed in the previous subsection to the M/UM case. Note that, if there is no match, by independence, the BP of p_n, defined by (6), can be shown to be the same as

$$\mathrm{E}(p_n) = \int \mathrm{logit}^{-1}(x_n^\top \beta + \sigma\xi)\phi(\xi)d\xi, \tag{8}$$

where ϕ is the pdf of $\mathcal{N}(0, 1)$. In view of (8), an empirical predictor of p_n, in this situation, is (8) with β, σ replaced by $\hat{\beta}, \hat{\sigma}$, respectively. Denote the latter predictor by $\hat{p}_{n,e}$ (where the subscript e standards for "empirical"). Thus, in the M/UM case, we compute both $\hat{p}_{n,\hat{I}}$ [see (7)] and $\hat{p}_{n,e}$. We then compare $|\hat{p}_{n,\hat{I}} - \bar{y}_n|$ with $|\hat{p}_{n,e} - \bar{y}_n|$. If the former is smaller, the CMLMP of p_n is $\hat{p}_{n,\hat{I}}$; otherwise, it is $\hat{p}_{n,e}$. Denote the CMLMP defined this way by $\hat{p}_n$. The following theorem shows that, even if there does not exist a true match between the random effect associated with the new observations and one of the random effects associated with the training data, the CMLMP is still consistent in estimating, or predicting, p_n. Note that here the consistency is in terms of estimating p_n, not in terms of estimating the index I; in fact, $\hat{I}$ is not a consistent estimator of I, even in the matched case. This is similar to the finding of [6]. The proof is given in the Online Supplement.

Theorem 2. *Under the conditions of Theorem 1, we have* $\hat{p}_n - p_n \xrightarrow{\Pr} 0$, *as* m, $\min(n_1, \ldots, n_m)$, *and* $n_{\mathrm{new}} \to \infty$. *In other words, the CMLMP is consistent.*

2.2. General case: Unit-level covariates

We now consider a general case in which x_{ij} may depend on both i and j. We call this case unit-level covariates. Following the notation in Section 1, we propose the following strategy for matching the index I to one of the indexes $i \in \{1, \ldots, m\}$, regardless of whether or not there is an actual match. Write $\bar{p}_n = (p_{n,1} + \cdots + p_{n,\mathrm{new}})/n_{\mathrm{new}}$, and $X_i = (x_{ij}^\top)_{1\le j\le n_i}$, $y_i = (y_{ij})_{1\le j\le n_i}$. Define

$$\psi(X_i, y_i, \beta, \alpha_i) = \sum_{j=1}^{n_i}\{\alpha_i y_{ij} - \ln(1 + e^{x_{ij}^\top\beta+\alpha_i})\} = \alpha_i y_{i\cdot} - \sum_{j=1}^{n_i}\ln(1 + e^{x_{ij}^\top\beta+\alpha_i}).$$

If $I = i$, then, it can be shown that the BP of $\bar{p}_n$ is given by

$$\tilde{\bar{p}}_{n,i} = \frac{1}{n_{\mathrm{new}}}\sum_{k=1}^{n_{\mathrm{new}}}\frac{\int \mathrm{logit}^{-1}(x_{n,k}^\top\beta + \sigma\xi)\exp\{\psi(X_i, y_i, \beta, \sigma\xi)\}\phi(\xi)d\xi}{\int \exp\{\psi(X_i, y_i, \beta, \sigma\xi)\}\phi(\xi)d\xi}, \tag{9}$$

where, again, ϕ denotes the pdf of $\mathcal{N}(0, 1)$. The EBP of $\bar{p}_n$, denoted by $\hat{\bar{p}}_{n,i}$, is obtained by (9) with β, σ replaced by $\hat{\beta}, \hat{\sigma}$, respectively. We then identify I by

$$\hat{I} = \underset{i\in\{1,\ldots,m\}}{\mathrm{argmin}}|\hat{\bar{p}}_{n,i} - \bar{y}_n|, \tag{10}$$

where $\bar{y}_n$ is defined the same as before.

Once $\hat{I}$ is obtained, the CMLMP of $p_{n,k}$, for any $k \in \{1, \ldots, n_{\mathrm{new}}\}$, is obtained as

$$\hat{p}_{n,k} = \left.\frac{\int \mathrm{logit}^{-1}(x_{n,k}^\top\hat{\beta} + \hat{\sigma}\xi)\exp\{\psi(X_i, y_i, \hat{\beta}, \hat{\sigma}\xi)\}\phi(\xi)d\xi}{\int \exp\{\psi(X_i, y_i, \hat{\beta}, \hat{\sigma}\xi)\}\phi(\xi)d\xi}\right|_{i=\hat{I}}. \tag{11}$$

Alternatively, the mixed effect $p_{n,k}$ can be predicted as

$$\check{p}_{n,k} = \mathrm{logit}^{-1}(x_{n,k}^\top\hat{\beta} + \hat{\alpha}_{\hat{I}}), \tag{12}$$

where $\hat{\alpha}_i$ is the EBP of α_i given by (4) with β, σ replaced by $\hat{\beta}, \hat{\sigma}$, respectively.

Similar to Theorems 1 and 2, consistency of the CMLMP defined by (11) or (12) can be established. Due to similarity of the results, and also to avoid making the paper too theoretical, the details are omitted. See the next subsection for empirical performance.

2.3. Empirical demonstration

To demonstrate the theoretical results established in the previous section and also to study empirical performance of CMLMP, we carry out a number of simulation studies. We first consider a "matched case", in which there is a match between I, the group number for the new observations, and one of the indexes $i \in \{1, \ldots, m\}$ in the training data; an "unmatched case", in which such a match does not exist, will be considered next. Throughout this section, the expectations in (3) and (4) are evaluated via Monte Carlo method.

Table 1
Simulated MSPEs of CMLMP and SLRP.

m	10	50	100	500	1000
CMLMP	0.0206	0.0162	0.0155	0.0196	0.0198
SLRP	0.0245	0.0231	0.0242	0.0238	0.0230
% Improve	19.17%	42.78%	56.25%	21.48%	16.19%

Table 2
Simulated MSPEs of CMLMP and SLRP.

n_{new}	1	5	10	100	1000
CMLMP	0.0361	0.0175	0.0130	0.0061	0.0041
SLRP	0.0225	0.0228	0.0247	0.0247	0.0232
% Improve	−37.72%	30.26%	89.22%	304.99%	465.93%

Table 3
Simulated MSPEs of CMLMP and SLRP.

σ	0.25	0.5	1	2	3
CMLMP	0.0038	0.0087	0.0158	0.0186	0.0171
SLRP	0.0025	0.0074	0.0232	0.0622	0.1038
% Improve	−34.39%	−15.44%	46.66%	234.89%	507.22%

2.3.1. The matched case

We consider a case where the covariates are at the cluster level. The training data are generated under a mixed logistic model defined, for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n_i\}$, by

$$\text{logit}(p_i) = \text{logit}\{\Pr(y_{ij} = 1|\alpha_i)\} = 1 + 2x_i + \alpha_i, \tag{13}$$

The covariates, x_i, are generated from the $\mathcal{N}(0, 1)$ distribution. The random effects, α_i, are then generated independently from the $\mathcal{N}(0, \sigma^2)$ distribution. The true index, I, is equal to 1 so that $\alpha_I = \alpha_1$. The x_n for the new observations, $y_{n,1}, \dots, y_{n,n_{\text{new}}}$, is also generated from $\mathcal{N}(0, 1)$; the new observations are then generated under a similar model as (13) with x_i, α_i replaced by x_n, α_I, respectively.

The results, based on 500 simulation runs, are reported in Tables 1–3. As a comparison, we also include the results of SLRP (see the last paragraph of Section 1). The results are for $n_1 = \cdots = n_m = 5$. Table 1 reports the results with $\sigma = 1$, $n_{\text{new}} = 5$, and changing m, the number of groups, or clusters, in the training data. Table 2 reports the results with $\sigma = 1$, $m = 50$, and changing n_{new}. Table 3 reports the results with $m = 50$, $n_{\text{new}} = 5$, and changing σ. Reported are simulated MSPEs (i.e., MSPEs based on the simulations) with the % Improve defined as

$$\text{\% Improve} = \left(\frac{\text{MSPE of SLRP} - \text{MSPE of CMLMP}}{\text{MSPE of CMLMP}}\right) \times 100\%.$$

It is seen that, as long as σ^2 and n_{new} are not too small, CMLMP improves over SLRP, but the improvement patterns are different as m, n_{new}, or σ increase. As n_{new} increases, so is the amount of improvement. This is because, as n_{new} increases, so is the accuracy of the class identification, which can be seen from (7); namely, $\bar{y}_n$ converges to $p_n = \text{logit}^{-1}(x_n^\top \beta + \alpha_I)$, which is the true mixed effect of interest. In contrast, as m increases, the amount of improvement of CMLMP does not always increase; in fact, the best improvement appears to take place when $m = 100$. One interpretation is that, as m increases, it is getting harder to correctly identify the class (because there are more choices to choose from). Although Theorem 2 shows consistency of CMLMP in spite of misidentification of the class (i.e., in the unmatched case), the result holds under the condition that $n_i \to \infty$ for every i. In the case of our simulation, $n_i = 5$ for all i. Finally, as σ increases, so does the relative improvement of CMLMP over SLRP. This makes sense because the larger σ, the more SLRP is missing by not taking into account the clustering in the population when making the prediction.

2.3.2. Comparing matched and unmatched cases

Following [6], we consider a situation where there is no match between the group of the new observations and a group among the training data, and compare this case with the case where there is a match. Specifically, we consider the following two scenarios. Scenario I (matched case): $\alpha_I = \alpha_1$. Scenario II (unmatched case): α_I is generated independently with $\alpha_1, \dots, \alpha_m$. The results, again based on 500 simulation runs, are reported in Tables 4–6, with the same set-ups as Tables 1–3. One thing that is new here is % Match, which is the percentage of times that the CMLMP procedure thinks that there is a match. Recall that, in Section 2.1.2, one compares $|\hat{p}_{n,\hat{I}} - \bar{y}_n|$ with $|\hat{p}_{n,e} - \bar{y}_n|$; if the former is smaller, CMLMP thinks that there is a match, and therefore uses $\hat{p}_{n,\hat{I}}$ as the predictor; otherwise, CMLMP thinks that there is no match, thus uses $\hat{p}_{n,e}$ as the predictor.

Table 4
Simulated MSPEs of CMLMP and SLRP.

Scenario	m	10	50	100	500	1000
I	CMLMP	0.0231	0.0167	0.0158	0.0187	0.0197
I	SLRP	0.0318	0.0247	0.0245	0.0233	0.0211
I	% Match	99.00%	99.80%	99.60%	99.80%	100.00%
II	CMLMP	0.0255	0.0181	0.0161	0.0177	0.0190
II	SLRP	0.0332	0.0216	0.0260	0.0249	0.0207
II	% Match	98.40%	99.40%	99.40%	100.00%	100.00%

Table 5
Simulated MSPEs of CMLMP and SLRP.

Scenario	n_{new}	1	5	10	100	1000
I	CMLMP	0.0330	0.0178	0.0109	0.0062	0.0050
I	SLRP	0.0249	0.0256	0.0232	0.0266	0.0218
I	% Match	100.00%	99.60%	99.20%	99.20%	99.00%
II	CMLMP	0.0375	0.0182	0.0104	0.0055	0.0047
II	SLRP	0.0244	0.0291	0.0237	0.0255	0.0219
II	% Match	100.00%	100.00%	99.20%	99.00%	99.00%

Table 6
Simulated MSPEs of CMLMP and SLRP.

Scenario	σ	0.25	0.5	1	2	3
I	CMLMP	0.0038	0.0085	0.0168	0.0176	0.0216
I	SLRP	0.0027	0.0079	0.0238	0.0673	0.1037
I	% Match	99.80%	99.80%	99.40%	99.60%	99.00%
II	CMLMP	0.0043	0.0101	0.0176	0.0209	0.0193
II	SLRP	0.0028	0.0091	0.0257	0.0658	0.1119
II	% Match	99.80%	99.60%	99.40%	98.20%	98.40%

It is seen that, regardless of the true matching status, CMLMP matches anyway, with % Match close to 100% under both scenarios. But, in spite of the mismatches under Scenario II (because there is supposed to be no match in this case), the performance of CMLMP is very similar under the two scenarios. In particular, CMLMP continues to improve over SLRP in spite of mismatching the class. This pattern is very similar to the finding of [6]. An interpretation is that, even if CMLMP does not match the exact class number, it matches a class that is very close to the true class in terms of prediction of the mixed effect of interest, which is all that matters. Furthermore, as long as m is reasonably large, there is always a random effect corresponding to the training data that almost matches the random effect corresponding to the new observations, even if there is no match; see the proof of Theorem 2 in the Online Supplement.

2.3.3. Unit-level covariate

We now consider a case that the covariate is at unit level. The data-generating model is the same as (13) except replacing x_i by x_{ij}, which are also generated from the $\mathcal{N}(0, 1)$ distribution. We consider the case with $m = 50$, $n_i = 5$, $n_{\text{new}} = 100$, $\sigma = 1$. An unmatched case is considered. The method of Section 2.2 is used to obtain the CMLMP, which is compared with SLRP. Instead of reporting the results in tables, we present the simulated MSPEs in Fig. 1. The results are based on 200 simulation runs. It is seen that CMLMP performs uniformly better for predicting the mixed effect associated with every one of the 100 new observations.

3. CMLMP incorporating cluster-level covariates

So far, our strategy for class identification is similar to [6] in that no information from the covariates of the new observations is used in identifying the class. This can be seen from (10), which only uses $\bar{y}_n$ from the new observations. In practice, covariate information can often help in identifying the class. This is particularly the case when there are covariates at the cluster level. Much effort has been made in trying to model the functional relationship between the mean response and the covariates, for example, linear regression, polynomial regression, splines, nonparametric regression. Alternatively, random effects are often introduced to “capture the uncaptured”, i.e., variation that cannot be explained by the assumed functional relationship with the covariates. In such a case, it is reasonable to assume that there is some kind of correspondence between the cluster-level covariates and the cluster-specific random effects, $\alpha_1, \ldots, \alpha_m$. This is something not considered by Jiang et al. [6] and a new contribution of the current paper.

Let w_i denote a vector of cluster-level covariates. Our idea is to consider the difference, $w_i - w_n$, where w_n is the corresponding vector of covariates associated with the new observations. The motivation is that, ideally, $w_n = w_i$ implies $I = i$; therefore, the difference between w_n and w_i should be taken into account in identifying the class. The question

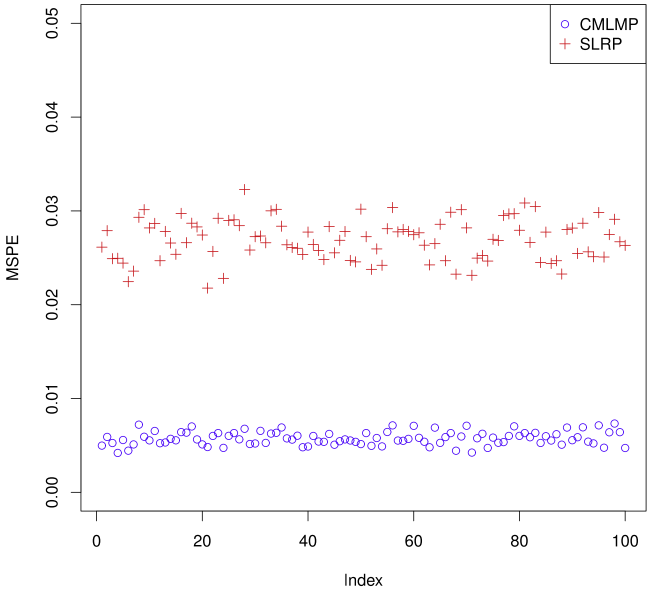


Fig. 1. Simulated MSPEs for prediction of mixed effects for 100 new observations.

is: How? Consider a cluster-specific linear predictor defined as $\ell_i = w_i^\top b + \alpha_i$, where b is the vector of fixed effects corresponding to w_i. Similarly, we have $\ell_n = w_n^\top b + \alpha_I$ for the new observations. Note that ℓ_i is part of the linear predictor on the right side of (1) that is cluster-specific; similarly, ℓ_n is part of the linear predictor on the right side of (5) corresponding to the new observations. However, in order to match the clusters, the parts ℓ_i and ℓ_n are all we need. We have $\ell_i - \ell_n = (w_i - w_n)^\top b + \alpha_i - \alpha_I$. Therefore, we have

$$\mathrm{E}(\ell_i - \ell_n)^2 = \{(w_i - w_n)^\top b\}^2 + \mathrm{E}(\alpha_i - \alpha_I)^2 = \mathrm{E}[\{(w_i - w_n)^\top b\}^2 + (\alpha_i - \alpha_I)^2].$$

Thus, using the CMMP idea [6], to minimize $\mathrm{E}(\ell_i - \ell_n)^2$ we need to minimize

$$\{(w_i - w_n)^\top b\}^2 + (\alpha_i - \alpha_I)^2. \tag{14}$$

Three things are unknown in (14): b, α_i, and α_I. Naturally, we replace b by $\tilde{b}$, the subvector of $\tilde{\beta}$, the consistent estimator of β based on the training data. It is also natural to replace α_i, α_I by their EBPs, $\tilde{\alpha}_i$, $\tilde{\alpha}_n$. The former is given by (4) with β and σ replaced $\tilde{\beta}$ and $\tilde{\sigma}$, the consistent estimator of σ. Similarly, the BP of α_I based on the new observations is

$$\tilde{\alpha}_n = \sigma \frac{\mathrm{E}[\xi \exp\{y_{n,\cdot}\sigma\xi - \sum_{j=1}^{n_{\rm new}} \ln(1 + e^{x_{n,j}^\top \beta + \sigma\xi})\}]}{\mathrm{E}[\exp\{y_{n,\cdot}\sigma\xi - \sum_{j=1}^{n_{\rm new}} \ln(1 + e^{x_{n,j}^\top \beta + \sigma\xi})\}]}. \tag{15}$$

The EBP $\tilde{\alpha}_n$ is obtained by replacing β, σ in (15) by $\tilde{\beta}$, $\tilde{\sigma}$, respectively. Our new criterion of class identification is, therefore, given by

$$\hat{I} = \operatorname*{argmin}_{i \in \{1,\ldots,m\}} [\{(w_i - w_n)^\top \tilde{b}\}^2 + (\tilde{\alpha}_i - \tilde{\alpha}_n)^2]. \tag{16}$$

Under regularity conditions, theoretical results similar to Theorems 1 and 2 can be established for CMLMP based on the new criterion. The detail is omitted. Below we present the results of a simulation study, in which we compare the performance of the new CMLMP procedure with SLRP and EBP without matching the class. In the latter case, the random effect α_I is assumed to be independent with $\alpha_1, \ldots, \alpha_m$; thus, the EBP is obtained by (3) with $y_{i\cdot}$, x_i, n_i replaced by $y_{n\cdot} = \sum_{k=1}^{n_{\rm new}} y_{n,k}$, x_n, $n_{\rm new}$, respectively, and β, σ replaced by $\hat{\beta}, \hat{\sigma}$, the consistent estimator of β, σ based on combination of the training data and new observations.

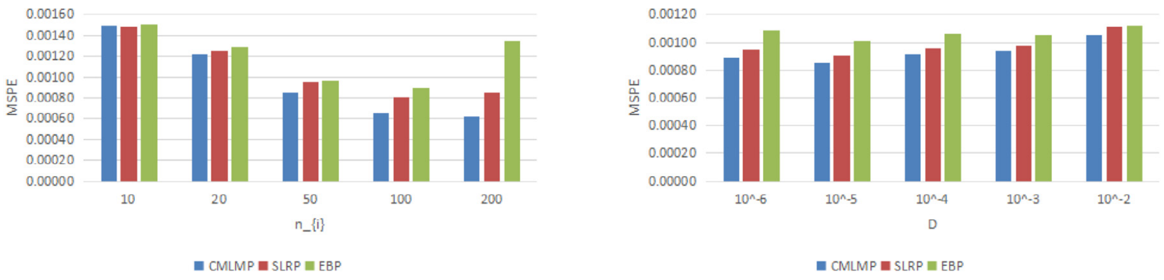
The simulation is under the same model, (13), considered in the previous section, with the cluster-level covariate $w_i = x_i$ except that the random effect is generated differently. Here, the random effect, α_i, is introduced to capture the uncaptured, as discussed at the beginning of this section. Specifically, $\alpha_i = g(w_i) + v_i$, where $g(w_i)$ corresponds to the uncaptured, an unknown function of the covariate, and v_i is a small noise, generated independently from the $\mathcal{N}(0, D)$ distribution. Two different functions are considered: $g(w_i) = w_i^3$; and $g(w_i) = w_i^2 - 1$. We consider a matched case with $I = 1$. The results, based on 500 simulation runs, are reported in Tables 7 and 8. Table 7 is for $m = 50$, $n_{\rm new} = 10$, $D = 10^{-4}$ and changing n_i;

Table 7
Simulated MSPEs of CMLMP, SLRP, EBP: $m = 50$, $n_{\text{new}} = 10$, $D = 10^{-4}$.

	n_i	10	20	50	100	200
$g(w_i) = w_i^3$	CMLMP	.00150	.00122	.00086	.00066	.00063
	SLRP	.00149	.00125	.00096	.00080	.00085
	EBP	.00151	.00129	.00097	.00090	.00135
$g(w_i) = w_i^2 - 1$	CMLMP	.01074	.00836	.00821	.00525	.00585
	SLRP	.02113	.02087	.02289	.01692	.02492
	EBP	.01035	.00998	.01083	.01034	.00980

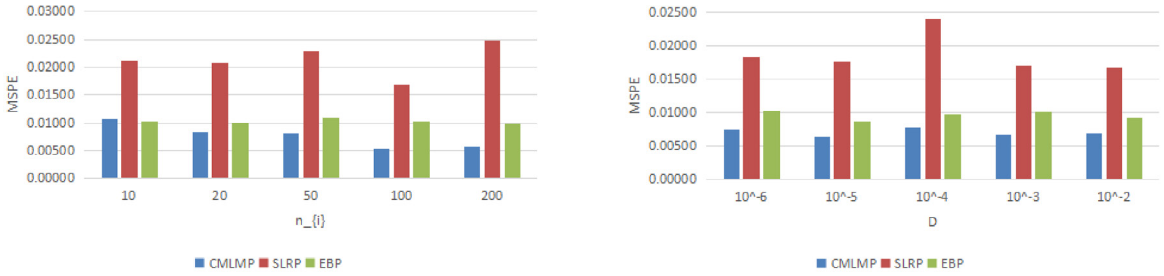
Table 8
Simulated MSPEs of CMLMP, SLRP, EBP: $m = 50$, $n_{\text{new}} = 10$, $n_i = 50$.

	D	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
$g(w_i) = w_i^3$	CMLMP	.00089	.00085	.00092	.00094	.00105
	SLRP	.00095	.00091	.00096	.00097	.00111
	EBP	.00109	.00101	.00106	.00105	.00112
$g(w_i) = w_i^2 - 1$	CMLMP	.00735	.00640	.00781	.00661	.00681
	SLRP	.01840	.01772	.02403	.01702	.01669
	EBP	.01028	.00859	.00973	.01003	.00913



(a) (b)

Fig. 2. Simulated MSPEs for $g(w_i) = w_i^3$: (a) changing n_i; (b) changing D; within each group of bar charts: blue (left)—CMLMP; red (middle)—SLRP; green (right)—EBP.



(a) (b)

Fig. 3. Simulated MSPEs for $g(w_i) = w_i^2 - 1$: (a) changing n_i; (b) changing D; within each group of bar charts: blue (left)—CMLMP; red (middle)—SLRP; green (right)—EBP.

Table 8 is for $m = 50$, $n_{\text{new}} = 10$, $n_i = 50$ and changing D. The corresponding bar charts [grouped according to different gs] are presented in Figs. 2 and 3.

It is seen that, as long as n_i is not too small (i.e., $n_i > 10$), CMLMP outperforms both SLRP and EBP. As for the degree of improvement, the relative improvement is much more significant for $g(w_i) = w_i^2 - 1$ than for $g(w_i) = w_i^3$. This can be seen clearly from Figs. 2 and 3. The advantage of CMLMP over SLRP and EBP increases as n_i increases; see part (a) of those figures. In contrast, the advantage of CMLMP over the other methods does not seem to change with D, ranging from 10^{-6} to 10^{-2}. Also, for $g(w_i) = w_i^2 - 1$, both CMLMP and EBP outperform SLRP; while for $g(w_i) = w_i^3$, only CMLMP outperforms SLRP (except for one case with $n_i = 10$).

4. Measures of uncertainty

In this section, we develop a procedure that produces a second-order unbiased estimator of the MSPE of CMLMP in the sense that

$$\mathrm{E}(\widehat{\mathrm{MSPE}}) = \mathrm{MSPE} + o(m^{-1}), \tag{17}$$

where $\widehat{\mathrm{MSPE}}$ denotes the estimator of MSPE. Let θ denote the mixed effect of interest, for example, $\theta = \mathrm{logit}^{-1}(x_{\mathrm{new},k}^{\top}\beta + \alpha_I)$, for a fixed $k \in \{1, \ldots, n_{\mathrm{new}}\}$; see above Eq. (5). Let $\hat{\theta}$ denote the CMLMP of θ. Also, let $\psi = (\beta, \sigma^2, I)$, where I is the true group index for the new observations. The MSPE of $\hat{\theta}$ can be expressed as

$$\mathrm{MSPE} = \mathrm{E}(\hat{\theta} - \theta)^2 = \mathrm{E}[\mathrm{E}\{(\hat{\theta} - \theta)^2|Y\}], \tag{18}$$

where $Y = (y, y_{\mathrm{new}})$, y denotes the training data, and y_{new} the new observations. The conditional expectation inside the outer expectation on the right side of (18) is a function of Y and ψ, i.e.,

$$a(Y, \psi) = \mathrm{E}\{(\hat{\theta} - \theta)^2|Y\} = \hat{\theta}^2 - 2\hat{\theta}\mathrm{E}(\theta|Y) + \mathrm{E}(\theta^2|Y) = \hat{\theta}^2 - 2\hat{\theta}a_1(Y, \psi) + a_2(Y, \psi), \tag{19}$$

where $a_s(Y, \psi) = \mathrm{E}(\theta^s|Y)$ for each $s \in \{1, 2\}$. If we replace the ψ in (19) by $\hat{\psi}$, a consistent estimator of ψ, the result is a first-order unbiased estimator, i.e.,

$$\mathrm{E}\{a(Y, \hat{\psi}) - a(Y, \psi)\} = O(m^{-1}). \tag{20}$$

Furthermore, both $\mathrm{MSPE} = \mathrm{E}\{a(Y, \psi)\}$ [by (18), (19)] and $\mathrm{E}\{a(Y, \hat{\psi})\}$ are functions of ψ. Let $b(\psi) = \mathrm{E}\{a(Y, \psi)\}$, $c(\psi) = \mathrm{E}\{a(Y, \hat{\psi})\}$, and $d(\psi) = b(\psi) - c(\psi)$. Then, (20) implies that $d(\psi) = O(m^{-1})$; thus, if we replace, again, ψ by $\hat{\psi}$ in $d(\psi)$, the difference is a lower-order term, viz.

$$d(\hat{\psi}) - d(\psi) = o(m^{-1}) \tag{21}$$

(in a suitable sense; e.g., in probability). Now consider the estimator

$$\widehat{\mathrm{MSPE}} = a(Y, \hat{\psi}) + b(\hat{\psi}) - c(\hat{\psi}). \tag{22}$$

We have, by combining (18)–(22), that

$$\mathrm{E}(\widehat{\mathrm{MSPE}}) = \mathrm{E}\{a(Y, \psi)\} + \mathrm{E}\{a(Y, \hat{\psi}) - a(Y, \psi)\} + \mathrm{E}\{d(\hat{\psi})\} = \mathrm{MSPE} + \mathrm{E}\{d(\hat{\psi}) - d(\psi)\} = \mathrm{MSPE} + o(m^{-1}),$$

which is (17). Essentially, this simple, heuristic derivation shows the second-order unbiasedness of the proposed MSPE estimator, (22), which can be made rigorous by imposing regularity conditions; see [5].

Note that, in (22), $a(Y, \hat{\psi})$ is the leading term which is typically $O(1)$; the remaining term, $d(\hat{\psi}) = b(\hat{\psi}) - c(\hat{\psi})$ is typically $O(m^{-1})$. However, the remaining term is usually much more difficult to evaluate than the leading term. Thus, we propose to evaluate this term via a Monte Carlo method.

Let P_{ψ} denote the distribution of Y with ψ being the true parameter vector. Given ψ, one can generate Y under P_{ψ}. For each $k \in \{1, \ldots, K\}$, let $Y_{[k]}$ denote Y generated under the kth Monte Carlo sample. Then, we have

$$b(\psi) - c(\psi) \approx \frac{1}{K}\sum_{k=1}^{K}\left\{a(Y_{[k]}, \psi) - a(Y_{[k]}, \hat{\psi}_{[k]})\right\}, \tag{23}$$

where $\hat{\psi}_{[k]}$ denotes $\hat{\psi}$ based on $Y_{[k]}$. If K is sufficiently large, which one has control over during the Monte Carlo simulation, the difference between the two sides of (23) is $o(m^{-1})$. Thus, we can replace the term $b(\hat{\psi}) - c(\hat{\psi})$ in (22) by the right side of (23) with ψ replaced by $\hat{\psi}$, leading to a Monte Carlo assisted (McA) MSPE estimator given by

$$\widehat{\mathrm{MSPE}}_K = a(Y, \hat{\psi}) + \frac{1}{K}\sum_{k=1}^{K}\left\{a(Y_{[k]}, \hat{\psi}) - a(Y_{[k]}, \hat{\psi}_{[k]})\right\}, \tag{24}$$

where $Y_{[1]}, \ldots, Y_{[K]}$ are generated as above with $\psi = \hat{\psi}$, and $\hat{\psi}_{[k]}$ is the estimator of ψ based on $Y_{[k]}$. As in [5], it can be shown that the McA MSPE estimator, (24), is second-order unbiased provided that $m/K \to 0$.

In order to apply the McA estimator to our case, we need to specify the function $a(Y, \psi)$, which, by (19), depends on $a_s(Y, \psi)$ with $s \in \{1, 2\}$. Consider a matched case. Note that, given $I \in \{1, \ldots, m\}$, we can combine the cluster in the training data corresponding to I with the new data, which are also associated with α_I. Let y_{Ik} with $k \in \{n_I + 1, \ldots, n_I + n_{\mathrm{new}}\}$ denote the expanded responses for the Ith cluster with $y_{Ik} = y_{\mathrm{new},k-n_I}$ for all $k \in \{n_I + 1, \ldots, n_I + n_{\mathrm{new}}\}$. Similarly, let $x_{Ik} = x_{\mathrm{new},k-n_I}$ for all $k \in \{n_I + 1, \ldots, n_I + n_{\mathrm{new}}\}$. Then, similar to Eq. (2), we have, for $s \in \{1, 2\}$,

$$a_s(Y, \psi) = \frac{\int \{h(x_{\mathrm{new},j}^{\top}\beta + \alpha_I)\}^s f_{\beta}(Y_{I\cdot}, X_I, \alpha_I) f_{\sigma^2}(\alpha_I) d\alpha_I}{\int f_{\beta}(Y_{I\cdot}, X_I, \alpha_I) f_{\sigma^2}(\alpha_I) d\alpha_I}, \tag{25}$$

for each $j \in \{1, \ldots, n_{\text{new}}\}$, where $h(u) = \text{logit}^{-1}(u)$, f_{σ^2} denotes the pdf of $\mathcal{N}(0, \sigma^2)$, and

$$f_\beta(Y_{I\cdot}, X_I, \alpha_I) = \exp\left\{ Y_{I\cdot}\alpha_I - \sum_{k=1}^{n_I+n_{\text{new}}} \ln(1 + e^{x_{Ik}^\top \beta + \alpha_I}) \right\}$$

with $Y_{I\cdot} = \sum_{k=1}^{n_I+n_{\text{new}}} y_{Ik}$ and $X_I = (x_{Ik}^\top)_{1\le k\le n_I+n_{\text{new}}}$. Expression (25) is evaluated by numerical integration [e.g., using the `integrate` function in `R`] or a Monte Carlo method.

Once $\widehat{\text{MSPE}}_K$ is obtained, we can use its square root as a measure of uncertainty for the CMLMP, and we do this for every mixed effect of interest. The method will be illustrated in the next section when we revisit the ECMO data.

5. Application: ECMO data revisited

We focus on the two outcomes of interest, bleed_binary variable and DischargeMortalit1Flag variable, that were mentioned in Section 1. The data includes 8601 patients data from 42 hospitals. The numbers of patients in different hospitals range from 3 to 487. We first use a forward–backward (F–B) BIC procedure to build a mixed logistic model; see, e.g., [1]. Namely, we use a forward selection based on logistic regression to add covariate variables, one by one, until 50% of the variables have been added; we then carry out a backward elimination to drop the variables that have been added, one by one, until all of the variables are dropped. This F–B process generates a sequence of (nested) models, to which the BIC procedure [8] is applied to select the optimal model.

The F–B BIC procedure leads to a subset of 12 patient-level covariates out of a total of more than 20 covariates. The same 12 covariates were selected for both outcome variables. Specifically, in the selected model, the probability of hemorrhage complication (or mortality) is associated with number of days during hospitalization (LengthOfStay), major surgery during hospitalization (MajSurgduringHosp_binary; Yes/No), whether the patient is no more than 30 days old (age_ind1), whether the patient has had at least one of the following: 747—Other congenital anomalies of circulatory system; 746—Other congenital anomalies of heart, excluding endocardial fibroelastosis; 745—Bulbus cordis anomalies and anomalies of cardiac septal closure; 770—Other respiratory conditions of fetus and newborn; 756—Other congenital musculoskeletal anomalies, excluding congenital myotonic chondrodystrophy (Top5PrincDx), whether the patient is flagged for cardiovascular (flag_CV; Yes/No), hematologic/immunology (flag_hemimm; Yes/No), metabolic (flag_metab; Yes/No), neuromuscular (flag_neuromusc; Yes/No), other congenital/genetic (flag_congengen; Yes/No), or respiratory (flag_resp; Yes/No), number of days under ECMO during hospitalization (ALLecmodays), and whether the patient has received the AT treatment (AT; Yes/No).

Out of the 12 patient-level covariates, two are continuous corresponding to number of days during hospitalization and the number of days under ECMO during hospitalization; the rest are binary. In addition to the patient-level covariates, there are two hospital-level covariates, namely, the total number of patients during the 10 year study who did receive *AT* (*yesat*) and total number of patients that were included in the 10-year study (*total*). Both hospital-level covariates are continuous. It should be noted that the four continuous covariates need to be standardized before carrying out the CMLMP analysis.

The proposed mixed logistic model includes the above 12 patient-level covariates as well as the two hospital-level covariates, plus a hospital-specific random effect that captures the "uncaptured" as well as between-hospital variation.

The mixed effects of interest are probabilities of hemorrhage complication corresponding to *bleed_binary*, and mortality probabilities associated with *DischargeMortalit1Flag*, for new observations. Note that, because most of the covariates are at the patient-level, these probabilities are patient-specific. However, the responses are clustered with the clusters corresponding to the hospitals, and there are 42 random effects associated with the hospitals under the mixed logistic model. In order to test the CMLMP method, we randomly select five patients from a given hospital and treat these as the new observations. The rest of the hospitals, and rest of the patients from the same hospital (if any), correspond to the training data. We then use the matching strategy of Section 3, with *yesat* and *total* as the cluster-level covariates that are used to identify the group for the new observations, then compute the CMLMP for each of the five selected patients. In addition, we compute the McA MSPE estimator of Section 4, given by (24) with $K = 100$, for each CMLMP, and use $\pm 2\{\widehat{\text{MSPE}}_K\}^{1/2}$ as the margin of error. This analysis applies to all but one hospital (Hospital #2033), for which only three patients are available. For this hospital all three patients are selected for the new observations, and the CMLMP and margin of error are obtained for all three patients. Therefore, for 41 out 42 of these analyses, there is a match between the new observations' group and one of the training data groups; and for one analysis there is no such a match. Overall, the analysis yields a total of 208 predicted probabilities with the corresponding margins of errors. The results are presented in Fig. 4 (bleed_binary) and Fig. A1 of the Online Supplement (DischargeMortalit1Flag). Note that, for DischargeMortalit1Flag, some of the predicted are close to zero; as a result, the lower margin is negative, and therefore truncated at 0. In contrast, all of the predicted probabilities are clearly above zero for the bleed_binary outcome; therefore, there is no need for truncation in terms of the margin of errors.

6. Concluding remarks

The main goal of CMLMP is prediction of subject-level mixed effects for new observations. The subject may correspond to a patient (such as in precision medicine), or a group of patients such as in the ECMO example. A key idea of CMLMP is matching the group. Although, in practice, the groups in the training data may be known, those for the new observations

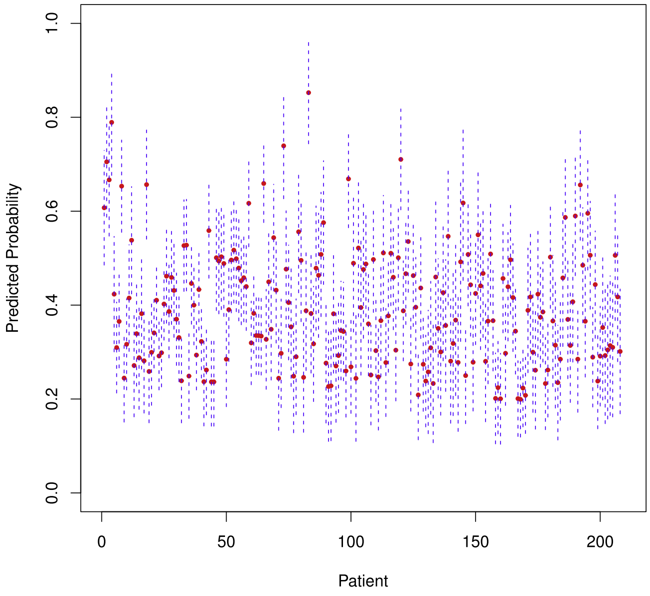


Fig. 4. Predicted probabilities of hemorrhage complication (bleed_binary) with margins of errors: Dash lines indicate margins of errors.

may be unknown. For example, if one wishes to use the ECMO data, which came from PHIS (see the middle of the second paragraph of Section 1), to predict the probability of hemorrhage complication for a new patient outside the PHIS, the group member for the new patient is unknown. One may argue, of course, that the new patient does not have a "group" corresponding to the training data, but the point is to make use of the rich source of information from the training data to improve prediction accuracy, and for that we need to match the new patient to an existing group in the training data in order to "borrow strength". Our theoretical and empirical results have shown that, even if there is no exact match between the new observation group and a group in the training data, CMLMP can still improve prediction accuracy. In fact, three of the 208 new observations in the above ECMO example do not have a match in the training data.

It should be noted that there are other situations, in practice, in which group memberships are hidden due to confidentiality concerns. For example, again in the ECMO example, the data were de-identified prior to inclusion in the study dataset (see the middle of the second paragraph of Section 1); luckily, encrypted medical records allowed tracking the patient's hospital identity. For a new patient who may wish to keep stringent confidentiality, even the hospital identity may not be known. This would create another situation that the group member of the new patient is unknown.

Although the group identification plays a key role in CMLMP, our main goal is prediction of the mixed effects, not identification of the group memberships. Our theoretical results have shown consistency in terms of prediction of the mixed effects, but not in terms of identifying the correct group membership. In fact, as m (the number of groups in the training data) increases, which is a key condition for consistency of the CMLMP of the mixed effect, the probability of identifying the correct group membership goes to 0 (rather than going to 1). But, in spite of that, our CMLMP of the mixed effect is consistent which is all that matters.

Acknowledgments

The research of Hanmei Sun and Yihui Luan was supported by the Natural Science Foundation of China Grants 11371227, 61432010 and 11626247. The research of Thuan Nguyen and Jiming Jiang was partially supported by the NSF grants DMS-1509557 and DMS-1512084, respectively. The authors wish to thank Dr. Trisha E. Wong of the Oregon Health & Science University for kindly allow them to use the ECMO data and for helpful discussions. Finally, the authors are grateful to an Associate Editor and two referees for their valuable comments that have led to improvement of the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2018.06.004.

References

[1] K.W. Broman, T.P. Speed, A model selection approach for the identification of quantitative trait loci in experimental crosses, J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (2002) 641–656.

[2] P. Glass, D.I. Bulas, A.E. Wagner, S.R. Rajasingham, L.A. Civitello, P.H. Papero, C.E. Coffman, B.L. Short, Severity of brain injury following neonatal extracorporeal membrane oxygenation and outcome at age 5 years, Dev. Med. Child Neurol. 39 (1997) 441–448.
[3] J. Jiang, Linear and Generalized Linear Mixed Models and their Applications, Springer, New York, 2007.
[4] J. Jiang, P. Lahiri, Empirical best prediction for small area inference with binary data, ann, Inst. Statist. Math. 53 (2001) 217–243.
[5] J. Jiang, P. Lahiri, T. Nguyen, A unified Monte Carlo jackknife for small area estimation after model selection, Ann. Math. Sci. Appl. (2018) in press.
[6] J. Jiang, J.S. Rao, J. Fan, T. Nguyen, Classified mixed model prediction, J. Amer. Statist. Assoc. 113 (2018) 269–279.
[7] W. Muntean, Fresh frozen plasma in the pediatric age group and in congenital coagulation factor deficiency, Thromb. Res. 107 (2002) S29–S32 [0049–3848 (Print); 0049–3848 (Linking)].
[8] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.