A beta-binomial mixed-effects model approach for analysing longitudinal discrete and bounded outcomes

Josu Najera-Zuloaga¹ | Dae-Jin Lee¹ | Inmaculada Arostegui¹,²,³

1Basque Center for Applied Mathematics, Bilbao, Spain
2Departamento de Matemática Aplicada y Estadística e Investigación Operativa, Universidad del País Vasco UPV/EHU, Bilbao, Spain
3Red de Investigación en Servicios de Salud en Enfermedades Crónicas, Bilbao, Spain

Correspondence
Dae-Jin Lee, Basque Center for Applied Mathematics, 48009 Bilbao, Spain.
Email: dlee@bcamath.org

Funding information
Ekonomiaren Garapen eta Lehiakortasun Saila, Eusko Jaurlaritzak, Grant/Award Number: ELKARTEK; Eusko Jaurlaritzak, Grant/Award Numbers: BERC 2014-2017, 2018–2021, IT-620-13; Secretaría de Estado de Investigación, Desarrollo e Innovación, Grant/Award Numbers: MTM2013-40941-P, MTM2014-52184-P, MTM2016-74931-P, MTM2017-82379-R, SEV-2013-0323; European Regional Development Fund, Grant/Award Number: RD12/0001/0001 REDISSEC (Red de Investigación en

Abstract
Patient-reported outcomes (PROs) are currently being increasingly used as primary outcome measures in observational and experimental studies since they inform clinicians and researchers about the health-status of patients and generate data to facilitate improved care. PROs usually appear as discrete and bounded with U, J, or inverse J shapes, and hence, exponential family members offer inadequate distributional fits. The beta-binomial distribution has been proposed in the literature to fit PROs. However, the fact that the beta-binomial distribution does not belong to the exponential family limits its applicability in the regression model context, and classical estimation approaches are not straightforward. Moreover, PROs are usually measured in a longitudinal framework in which individuals are followed up for a certain period. Hence, each individual obtains several scores of the PRO over time, which leads to the repeated measures and defines the correlation structure in the data. In this work, we have developed and proposed an estimation procedure for the analysis of correlated discrete and bounded outcomes, particularly PROs, by a beta-binomial mixed-effects model. Additionally, we have implemented the methodology in the PROreg package in R. Because there are similar approaches in the literature to address the same issue, this work also incorporates a comparison study between our proposal and alternative methodologies commonly implemented in R and shows the superior performance of our estimation procedure. This paper was motivated by the analysis of the health-status of patients with chronic obstructive pulmonary disease, where the main objective is the assessment of risk factors that may affect the evolution of the disease. The application of the proposed approach in the study leads to clinically relevant results.

KEYWORDS
beta-binomial distribution, mixed-effects models, patient-reported outcomes, PROreg R-package

1 | INTRODUCTION

Grouped binary observations are usually expected to conform to a binomial distribution. However, in the medical or biological fields grouped binary observations often appear as correlated measurements with skewed distributions, showing U, J, or inverse J shapes. In these cases, the resulting empirical variance might be larger than that specified by a binomial model, thus leading to overdispersion.

In this work, we are interested in self-reported outcomes, which usually have discrete and bounded distributions. A particular case of self-reported outcomes in medicine are the patient-reported outcomes (PROs). PROs are often built as a sum of ordinal or
binary responses to several items from the same questionnaire that have discrete and bounded responses, and they usually display dispersed forms and accumulate values in one or both edges of the distribution scale (Arostegui, Núñez-Antón, & Quintana, 2012). The different items are answered by the same individual. Therefore, there is a correlation structure within the outcomes provided by the questionnaire. In fact, correlation invalidates the binomial distribution which assumes that the binary responses are independent. Arostegui, Núñez-Antón, and Quintana (2007) proposed the beta-binomial distribution as an adequate fit to some particular PROs. The authors showed the adequacy of the beta-binomial distribution compared to other commonly used exponential family members, such as binomial and normal distributions, in a cross-sectional framework.

The beta-binomial model is defined as a mixture between two conjugate distributions. A distribution is said to be a natural conjugate to a given sampling process if its probability density (or mass) function is proportional to a likelihood function corresponding to some conceivable sample from the process (see Dickey, 1982; Bernardo & Smith, 1994, for further details). Indeed, the beta-binomial model takes into account the correlation structure within the grouped binary observations, and it also allows for a flexible relationship between the mean and the variance of the outcome.

In a regression framework with independent observations, there are two main approaches in the literature for addressing mixed distributions: (a) the marginal approach and (b) the conditional approach. On the one hand, the marginal approach applies a predefined link function to the marginal expectation of the distribution and connects it with the given covariates. The estimation procedure is maximum likelihood since the conjugacy allows for a closed-form marginal density function. On the other hand, the conditional approach applies a link function to the conditional expectation of the model and includes the conjugate random effects in the linear predictor. In this case, the estimation of the model occurs in a mixed-effects framework where some marginal likelihood approximations are considered.

The fact that the beta-binomial distribution does not belong to the exponential family limits its extension to hierarchical structures, such as longitudinal or clustered data. Indeed, the estimation of fixed and random effects and variance components in a mixed-effects framework from the responses drawn from nonexponential family distributions are not common in the literature. In the particular case of the beta-binomial regression, one can consider two approaches to construct a mixed-effects model: the marginal and the conditional models (see Najera-Zuloaga, Lee, & Arostegui, 2017). Based on the conditional approach Lee and Nelder (1996) developed the hierarchical generalised linear models (HGLMs) in which additional conjugate random effects can be included in the linear predictor of a classical generalised linear model (GLM; McCullagh & Nelder, 1989) as mixed-effects model (GLMMs, McCulloch & Searle, 2001). Moreover, the R software implementation of this methodology is available through the hglm package by Ronnegard, Shen, and Alam (2010).

Another extension of the conditional approach for hierarchical data is the combined models developed by Molenberghs, Verbeke, Demétrio, and Vieira (2010); Molenberghs, Verbeke, Iddi, and Demétrio (2012). Combined models are a broad class of models that include nonlinear mixed-effects models that multiply the conjugate random effects with a nonlinear transformation of the conditional mean. The beta-binomial combined model is developed in George, Iddi, and Molenberghs (2016), and the software implementation is performed by the NLMIXED procedure in SAS for nonlinear mixed effects. Similar nonlinear mixed-effects models are available in R, in the nlmixed package (Pinheiro, Bates, Sarkar, & R Core Team, 2016). Wu (2009) details different numerical approximations in mixed-effects models available in statistical software.

Because the beta and binomial are conjugate distributions, both the HGLM and combined model approaches could be applied in the hierarchical PRO's framework. The model's definition would be similar in both methodologies since they assume that, conditioned on Gaussian and beta (conjugate) random effects, the outcome follows a binomial distribution. Hence, they attempt to model the probability parameter of the conditional binomial distribution by means of a logit link function. Both approaches consider a logistic regression with Gaussian random effects. However, they are considerably different in the way in which the conjugate random effects are included. While in the HGLM approach a logit transformation of the random effects is additively included in the linear predictor, in the combined models, the random effects do not follow a linear structure. In fact, the combined models multiply the conjugate random effects with the antilogit transformation of the linear predictor of the model. Nevertheless, they share the same features that could make them inconvenient in a longitudinal PROs framework. First, the inclusion of additional conjugate random effects in GLMMs, both in a multiplicative or an additive way, can mask the useful interpretation that the Gaussian random effects offer in terms of individual differences at the baseline and the evolution in longitudinal models. Moreover, Najera-Zuloaga et al. (2017) showed that, in a regression framework with independent observations, the conditional approach is not always able to assess the statistical significance of the effect of the covariates in the outcome. Furthermore, the combined model for the beta-binomial distribution fails to generate a U-shaped distribution (Molenberghs et al., 2012), which is quite common in PRO studies (Arostegui et al., 2012).

Consequently, we will focus on the extension of the marginal beta-binomial regression approach to hierarchical structures, such as longitudinal or clustered data. The main objective of this work is the development of an estimation procedure for mixed-effects models with a beta-binomial outcome. We propose the development of a beta-binomial mixed-effects model that includes
Gaussian random effects in the linear predictor of a marginal beta-binomial regression model, which we call the $BBnm$ model. In particular, our proposal will provide an adequate methodology for analysing the correlated binary grouped outcomes in the hierarchical framework, especially in longitudinal PRO studies. In addition, we provide the researchers with an R package, PROreg, which is available at CRAN, to implement the proposed methodology.

There is an alternative methodology named the generalised additive models for location, scale, and shape (GAMLSS), which was developed by Rigby and Stasinopoulos (2005). It models the nonexponential family of distributions that include the marginal beta-binomial mixed-effects model as a particular case. Although the $BBnm$ approach falls under the GAMLSS framework, the estimation procedure we propose is different from the one used by GAMLSS. Our proposal distinguishes between the non-canonical parameters in the joint estimation process in a similar fashion as Lee and Nelder (2001) proposed for the HGLM methodology, whereas GAMLSS does not. Therefore, the second aim of this work is the comparison between our estimation procedure and GAMLSS. We will show that the estimates computed using the methodology that we propose are superior to the GAMLSS estimates in terms of bias and coverage.

The paper is organised as follows. In Section 2, we introduce the COPD study, which consists of temporally measured PROs. Indeed, this study has been the motivation of our work, due to the statistical difficulties we came across when trying to analyse this type of data. The model and the estimation proposal are detailed in Section 3, where the description and comparison of similar approaches is also discussed. In Section 4, we perform a simulation study that reveals the good properties of our proposal and its improvement compared to the GAMLSS approach. In Section 5, we present the analysis of the COPD study, where clinically and statistically significant results are obtained. Finally, Section 6 provides some conclusions and final remarks.

## 2 | MOTIVATION DATA

COPD is one of the most prevalent chronic diseases in the world, especially in developed countries (Buist, Vollmer, & McBurnie, 2008). Furthermore, its prevalence is expected to increase over the coming years since it is one of the major causes of mortality, and it is associated with a high risk of disability (Murray & López, 1997). Although the clinical assessment of COPD involves lung function parameters, such as FEV1% and dyspnoea, the overall impact of the disease on subjects is multifaceted. Currently, the progression of the disease is measured jointly with physical parameters and subjective PROs, such as health-status or Health-Related Quality of Life (HRQoL) questionnaires (Wiklund, 2004).

The COPD study is an observational study that was designed at the Respiratory Unit at Galdakao Hospital in Spain. Patients' COPD had to be stable for at least 6 months before enrolment, and they were followed up to 5 years with clinical examination and interviews. The health-status was measured with both a generic and a specific questionnaire, namely, the SF-36 Health Survey (Ware, Snow, Kosinski, & Gandek, 1993) and the St. George’s Respiratory Questionnaire (SGRQ) (Jones, Quirk, & Baveystock, 1991). More detailed information on the COPD study can be found in Esteban et al. (2016). One of the main objectives of the COPD study was to describe the HRQoL and the evolution of patients with COPD when they were stable. Moreover, researchers were interested in the variables that were related to the HRQoL or the evolution of these patients. Therefore, our aim was to detect sociodemographic and clinical variables that could significantly affect the HRQoL or the evolution of the HRQoL in patients with stable COPD. Previously, we used a beta-binomial regression approach in order to detect the variables that significantly affected HRQoL in a cross-sectional framework (Najera-Zuloaga et al., 2017). However, in this work, we are interested in the detection of variables that significantly affect the evolution of the HRQoL, and hence a longitudinal model is proposed.

The SF-36 Health Survey (version 1.2) which corresponds to the version 1.4 of the Spanish version, was selected as a generic HRQoL instrument. The validity and reliability of this instrument have been broadly tested (Stansfeld, Roberts, & Foot, 1997). Indeed, the SF-36 Health Survey was rated in 2002 by the British Medical Journal (Garratt, Schmidt, Mackintosh, & Fitzpatrick, 2002) as the most frequently used PRO of generic health in scientific publications. The SF-36 questionnaire has 36 items with different answer options. It was constructed to represent eight health dimensions, which are the physical functioning, role physical, bodily pain, general health, vitality, social functioning, role emotional, and mental health. The standardised scoring system was thoroughly described by the original authors (Ware et al., 1993). The raw dimensions of the SF-36 are usually standardised to a 0–100 scale, where a higher value means a better health-status. However, due to the nature of the construction in which ordinal or binary items are summed up, the dimensions can only reach some specific values on the [0,100] scale. Therefore, Arostegui, Núñez-Antón, and Quintana (2013) proposed and evaluated a method of recoding continuous and bounded scores, such as SF-36 scores, to a binomial form. The method is mainly based on the possible number of values that each dimension can obtain. A more detailed explanation of the recoding to a binomial form in the COPD study is detailed in Najera-Zuloaga et al. (2017).
Table 1: Descriptive evolution of the subtypes defined in Esteban et al. (2016) in the COPD study. The frequency and column percentage are shown for each subtype.

| Subtypes | Time framework | No. individuals | 1 Year | 2 Years | 5 Years |
|----------|----------------|-----------------|--------|---------|---------|
|          | Baseline       | n = 543         | 157    | 148     | 137     |
|          |                 | (32.70%)        | (34.82%)| (42.28%)|         |
|          | 1 Year         | n = 480         |        |         |         |
|          |                 | (35.91%)        |        |         |         |
|          | 2 Years        | n = 425         |        |         |         |
|          |                 | (36.47%)        |        |         |         |
|          | 5 Years        | n = 324         |        |         |         |
|          |                 | (35.18%)        |        |         |         |

Nonetheless, for the simplicity, clarity, and brevity of the exposition, we only show the results of three of the eight dimensions of the SF-36, namely, the role physical, general health, and role emotional. With respect to the covariates, Esteban et al. (2016) divided the subjects into the COPD study in four subtypes or clusters. Clusters were created based on multiple correspondence analysis combined with the automatic classification, including the analysis of most of the clinical variables collected in the study. The four subtypes showed different values for the variables that reflected the health-status of the subject. Three of the subtypes had a marked respiratory profile that showed a continuum in the severity, whereas the fourth subtype had a more systemic profile with intermediate respiratory conditions but a high prevalence of comorbidities. We will use this classification of subtypes as the independent variable instead of all the significant covariates. Table 1 shows a summary of the subjects and their subtypes during the follow-up.

Therefore, we will apply the BBmm approach to the longitudinal scores provided by the SF-36 in the COPD data. We will include the patient subtype as the covariate in the model in which the main goal will be the assessment of the evolution of the health-status of patients with COPD classified by subtype. In fact, we will be interested in analysing the differences in the HRQoL among the subtypes, either at the baseline or the evolution over time.

3. Beta-Binomial Mixed-Effects Model

In this section, we extend the marginal beta-binomial regression by allowing for the inclusion of random effects in the linear predictor. Let \( y' = (y_1, \ldots, y_n) \) be the vector of outcome variables and, conditional on the random effects \( u \), assume that the elements of \( y \) are independent and drawn from a beta-binomial distribution (see Appendix). To complete the model specification, assume that \( u \) follows a multivariate normal distribution with zero mean and variance-covariance matrix \( D \), which depends on a vector of variance parameters \( \lambda \) and is nonsingular. Hence, we have that

\[
y_i | u \sim BB(m_i, p_i, \phi) \quad \text{and} \quad u \sim N(0, D(\lambda)), \quad i = 1, \ldots, n. \tag{1}
\]

Notice that the specification of this beta-binomial mixed-effects model differs from the cited HGLM and the combined model methodologies since we assume that, conditional on the random effects, the outcomes follow a marginal beta-binomial distribution.

Let us define the parameter vector \( \theta = (\phi, \lambda) \), that consists of all the dispersion or variance components of the model. Following the marginal beta-binomial regression approach, we connect the probability parameter of the beta-binomial distribution with some given covariates \( X_1, \ldots, X_r \) and the random effects that are assumed to be fixed when conditioning by means of a logistic link function. That is,

\[
\eta_i = \log \frac{p_i}{1 - p_i} = x'_i \beta + z'_i u, \quad i = 1, \ldots, n, \tag{2}
\]

where \( p_i \) is the probability parameter of the beta-binomial distribution for the \( i \)th measurement, \( \beta' = (\beta_0, \beta_1, \ldots, \beta_r) \) is the vector of the fixed effects, \( x'_i \) is the \( i \)th row of a full rank \( n \times (r + 1) \) matrix \( X \) composed by the given covariates, \( u \) is a vector of length \( q \) composed by the random effects and \( z_i \) is the \( i \)th row of the \( n \times q \) model matrix \( Z \) composed by the random structure of the model.
The marginal likelihood for the model defined in Equation (2) is given by the following formula,

\[ L(\beta, \theta | y) = \int_{\mathbb{R}^q} \prod_{i=1}^n f(y_i | \beta, \phi, u) f(u | \lambda) du, \]

where \( q \) is the number of levels or components of the random effects \( u \), \( f(y_i | \beta, \phi, u) \) is the beta-binomial density function defined in Equation (A.1) and \( f(u | \lambda) \) is the multivariate normal density function for the random effects. There is no closed form for the marginal likelihood and, moreover, due to the complexity of the beta-binomial distribution, its numerical computation is almost intractable. In addition, it is totally uninformative about the realisations of the random effects. Consequently, approximation procedures must be developed in order to perform the estimation of the parameters in the model.

Equivalently, we can consider the marginal likelihood of the model in Equation (2) in an exponential form as

\[ L(\beta, \theta | y) = \int_{\mathbb{R}^q} \exp \left\{ \sum_{i=1}^n \log f(y_i | \beta, \phi, u) + \log f(u | \lambda) \right\} du. \]

Thus, considering the fact that the summation of twice differentiable regular functions is a twice differentiable regular function, we get the approximation

\[ \log L(\beta, \theta | y) \approx l(\beta, \theta | y, \tilde{u}) = h(\beta, \theta | y, \tilde{u}) - \frac{1}{2} \log \{ \det(M) \}, \tag{3} \]

where we skip constants and where

\[ h(\beta, \theta | y, u) = \sum_{i=1}^n \log f(y_i | \beta, \phi, u) + \log f(u | \lambda) \]

\[ = \sum_{i=1}^n \left[ \sum_{k=0}^{y_i-1} \log(p_i + k\phi) + \sum_{k=0}^{m_i-y_i-1} \log(1 - p_i + k\phi) - \sum_{k=0}^{m_i-1} \log(1 + k\phi) \right] - \frac{1}{2} \log \{ \det(D) \} - \frac{1}{2} u'D^{-1}u, \tag{4} \]

is the joint log-likelihood of the model, \( \beta \) and \( u \) enter in the formula through the linear predictor defined in Equation (2), \( \tilde{u} \) is the solution of \( \partial h / \partial u = 0 \) and \( -\frac{1}{2} \log \{ \det(M) \} \) is the adjusted term with

\[ M = \frac{\partial^2 h}{\partial u \partial u'} |_{u=\tilde{u}}. \]

The resulting approximation of the marginal log-likelihood is the first-order Laplace approximation, and it is equivalent to integrating the random effects out (Lee & Nelder, 2001).

### 3.1 Joint estimation of the fixed and random effects

For the estimation of the fixed effects, we assume that the dispersion parameter vector \( \theta \) is fixed and try to maximise the approximated log-likelihood of the model defined in Equation (3) which we denote as

\[ l(\beta | y, \tilde{u}, \theta) = A(\beta) + h(\beta | y, \tilde{u}, \theta). \]

Hence, the score equation of the fixed parameters of the model are given by

\[ S(\beta) = \frac{\partial A(\beta)}{\partial \beta} + \frac{\partial h(\beta | y, \tilde{u}, \theta)}{\partial \beta}. \tag{5} \]

In GLMMs, Breslow and Clayton (1993) showed that \( A(\beta) \) depends on \( \beta \) through the variance or weight matrix of the GLM working vector. Assuming that this variance or the weight matrix varies slowly as a function of \( \beta \), they proposed ignoring the term \( \partial A(\beta) / \partial \beta \) when obtaining the marginal maximum likelihood estimates.

Given that the beta-binomial distribution does not belong to the exponential family, the normalisation of the distribution through the working vector theory cannot be applied directly, and hence, the previous statement is not applicable in this case.
However, in the beta-binomial mixed-effects model, we have that
\[
A(\beta) = -\frac{1}{2} \log \{ \det(M) \} = -\frac{1}{2} \log \{ \det(Z'SWZ - D^{-1}) \},
\]
where \( S = \text{diag}(p_i(1 - p_i)), \ W = \text{diag}(w_i), \ u_i = -v_i p_i(1 - p_i) + \xi_i(1 - 2p_i) \) and
\[
\begin{cases}
\xi_i = \sum_{k=0}^{y_i-1} \frac{1}{p_i + k\phi} - \sum_{k=0}^{m_i-y_i-1} \frac{1}{1 - p_i + k\phi} \\
u_i = \sum_{k=0}^{y_i-1} \frac{1}{(p_i + k\phi)^2} + \sum_{k=0}^{m_i-y_i-1} \frac{1}{(1 - p_i + k\phi)^2}
\end{cases}
\]
for \( i = 1, \ldots, n \), with all the previous formulae evaluated at \( \mathbf{u} = \hat{\mathbf{u}} \). Hence, we have shown that \( A(\cdot) \) only depends on \( \beta \) through the weight matrices \( \mathbf{W} \) and \( \mathbf{S} \). Following Breslow and Clayton (1993), we assume that the weight matrices \( \mathbf{W} \) and \( \mathbf{S} \) vary slowly (or not at all) as a function of the fixed effects. Then, we ignore the first term in Equation (5), and maximise the second term to get the maximum likelihood estimates of the fixed effects.

We have shown that the adjusted term in the approximated log-likelihood in Equation (3) does not carry (almost) any information about the fixed effects, and hence, all the information regarding \( \beta \) is collected by the joint log-likelihood defined in Equation (4). Therefore, the adjusted term does not depend on \( \beta \) and hence, the random effects are canonical for the fixed effects. Lee, Nelder, and Pawitan (2006) showed that if the random effects in the model are canonical for the fixed effects, then the maximum likelihood estimator of the fixed effects from the marginal log-likelihood coincides with the joint maximiser of the joint log-likelihood. Therefore, we can derive the maximum likelihood estimator of the fixed effects by the maximisation of the joint log-likelihood with respect to the fixed and random effects.

The differentiation of the joint log-likelihood with respect to \( \beta \) and \( \mathbf{u} \) leads to the next score equations,
\[
\begin{cases}
\xi'\mathbf{S}\mathbf{X} = 0 \\
\xi'\mathbf{S}\mathbf{Z} - \mathbf{u}'\mathbf{D}^{-1} = 0
\end{cases}
\]
where \( \xi' = (\xi_1, \ldots, \xi_n), \xi_i, \ i = 1, \ldots, n \), and \( \mathbf{S} \) have been defined previously in Equation (6).

Different numerical algorithms can be used to solve the previous equations iteratively. However, due to the complexity of the second derivative of the beta-binomial density function, we propose using the delta method (Jørgensen, 1984), which is a modification of the Newton–Raphson method. This estimation method for the fixed and random effects is implemented in the \texttt{BBmm} function in the \texttt{PRDreg} package in R.

### 3.2 Estimation of variance components

For the estimation of the dispersion components \( \theta \) the maximum likelihood estimation might be substantially biased due to the previous estimation of \( \beta \) (Lee & Nelder, 2001). Many authors have defined different likelihood adjustments to perform the estimation of the variance components in several situations in the literature. For example, Patterson and Thompson (1971) developed a restricted or residual maximum likelihood (REML) criterion in linear mixed models (LMM) and Breslow and Clayton (1993) extended the approach to GLMMs. Lee and Nelder (1996) proposed a more general approach for the estimation of the dispersion components, which was called the adjusted profile \( h \)-likelihood. In fact, they showed that their proposal is a generalisation of the previously defined dispersion parameter estimation procedures. The adjusted profile \( h \)-likelihood is defined as
\[
h_p = h_A|_{\beta = \hat{\beta}, \mathbf{u} = \hat{\mathbf{u}}},
\]
\( h_A \) is the adjusted \( h \)-likelihood defined as
\[
h_A = h + \frac{1}{2} \log \{ \det(2\pi \mathbf{H}^{-1}) \},
\]
where \( h \) is the joint log-likelihood (see Equation (4)) and \( \mathbf{H} \) is the corresponding Hessian matrix of the model. The performed penalisation on the joint log-likelihood is equivalent to integrating the random effects out, as in the first-order Laplace’s approximation in Equation (3), and then, eliminating the nuisance fixed effects \( \beta \) by conditioning on the maximum likelihood estimates \( \hat{\beta} \). Lee and Nelder (1996) showed that maximum adjusted profile \( h \)-likelihood estimation can be derived for variance parameters by iteratively solving
\[
\frac{\partial h_A}{\partial \theta}|_{\beta = \hat{\beta}, \mathbf{u} = \hat{\mathbf{u}}} = 0,
\]
where \( \hat{\beta} \) and \( \hat{\mathbf{u}} \) are evaluated in each iteration.
Following the approximation of the second derivatives of the log-likelihood that the delta method provides, the Hessian matrix of the beta-binomial mixed-effects model is defined as

\[ H = \begin{pmatrix} X'SVX & X'SVZ \\ Z'SVX & Z'SVZ + D^{-1} \end{pmatrix}, \]

where \( V = \text{diag}(v_1, \ldots, v_s) \), with \( v_j \) and \( S \) were previously defined in Equation (6). It is worth noting that \( D \) is the only term in the Hessian matrix that depends on \( \lambda \). However, unlike the usual GLMM, or even the HGLM, where the dispersion parameter of the conditioned distribution can be explicitly removed from the Hessian matrix, in the beta-binomial mixed model the matrix \( V \) depends implicitly on the dispersion parameter \( \phi \). Accordingly, the computation of the score equation for the dispersion parameter of the beta-binomial distribution is computationally more expensive than in other models where the conditioned distribution belongs to the exponential family.

Hence, the score equations for the variance parameters of the presented beta-binomial mixed-effects model are defined as

\[
\frac{\partial h_\beta}{\partial \theta_i} = \frac{\partial h(\beta,\theta,u)}{\partial \theta_i}_{\beta=\beta, u=u} + \frac{1}{2} \frac{\partial \log(\det(H^{-1}))}{\partial \theta_i}, \quad i = 1, \ldots, k + 1,
\]

where \( k \) is the number of parameters needed for specifying \( D \), which is the length of the vector \( \lambda \). Notice that \( \theta_i \) refers to either the beta-binomial dispersion parameter \( \phi \) or a parameter of the vector of variance components \( \lambda \). In addition, because the variance components must be positive, it is common to rewrite the log-likelihood function and conduct the estimation for \( \log(\theta_i), i = 1, \ldots, k + 1 \).

Therefore, the iterative estimation algorithm of the BBmm approach can be summarised as follows.

- **i)** Give the initial values for \( \beta, u \) and \( \theta \).
- **ii)** For a fixed \( \theta \) estimate the random and fixed parameters by solving Equation (5).
- **iii)** For the previous estimates of \( \beta \) and \( u \), get the variance parameters by solving Equation (7).
- **iv)** Iterate between **ii)** and **iii)** until convergence occurs.

### 3.3 Alternative approaches in the literature

We have mentioned that the BBmm approach is based on the extension of the marginal beta-binomial regression model to a mixed-effects framework. Nevertheless, there are other methodologies in the literature that extend the marginal beta-binomial regression approach for the analysis of correlated data.

On the one hand, Wu, Zhang, and Long (2017) proposed a longitudinal beta-binomial model for overdispersed binomial data and estimated the regression parameters under a probit model using the generalised estimating equation (GEE) approach (Zeger & Liang, 1986). GEE is often used as a very general and computationally convenient alternative to mixed-effects regression models since it extends the classical GLMs to the case of correlated data. Although GEE is computationally easier than the mixed-effects approach, it is more restrictive regarding the drop out or incomplete follow-up assumptions, which can make the estimated mean responses quite different at the end of the study if future observations are related to the measurements that were made during the course of the study (Gibbons, Hedeker, & DuToit, 2010). Furthermore, mixed-effects models provide subject-specific effects that are quite useful both in longitudinal responses for the same subject and in individuals grouped in hierarchies.

On the other hand, the GAMLSS approach, developed by Rigby and Stasinopoulos (2005), is a very flexible methodology, since it addresses a very wide range of distributions and also allows for the inclusion of random effects to accommodate the correlation of the data among all the parameters of the given distribution. The GAMLSS approach models the vector of all the parameters \( \zeta' = (\zeta_1, \ldots, \zeta_s) \) of a general population probability density function \( f(y|\zeta) \). It assumes that, for \( j = 1, \ldots, s \), each of the distribution parameters \( \zeta_j \) is connected to some given covariates and random effects and that, given those random effects, the observations \( y_j, i = 1, \ldots, n \), are independent.

Let \( y' = (y_1, \ldots, y_n) \) be the set of outcome variables and \( g_j(\cdot) \) be a known monotonic link function connecting \( \zeta_j' = (\zeta_{j1}, \ldots, \zeta_{jn}) \) to the given covariates and random effects through

\[
g_j(\zeta_j) = \eta_j = X_j\beta_j + Z_ju_j,
\]

where \( \zeta_j \) and \( \eta_j \) are vectors of length \( n \), \( \beta_j' = (\beta_{0j}, \ldots, \beta_{r_jj}) \) is the fixed effects parameter vector, \( X_j \) is a known design matrix of order \( n \times (r_j + 1) \) composed by the given covariates, \( Z_j \) is a fixed known \( n \times q_j \) design matrix composed by the random...
structure of the model, and \( u_j \) is a random vector of length \( q_j \), for \( j = 1, \ldots, s \). The model assumes that the random effects \( u_j \) have a multivariate normal distribution, a mean of 0, and variance–covariance matrix \( D_j \), where \( D_j \) depends on a vector of hyperparameters \( \lambda_j \), for each \( j = 1, \ldots, s \).

The beta-binomial mixed-effects regression model displayed in Equation (2) reduces to the following definition of the GAMLSS model

\[
\begin{align*}
    g_1(\zeta_1) &= \eta_1 = X\beta_1 + Zu \\
    g_2(\zeta_2) &= \eta_2 = \beta_2.
\end{align*}
\]

In Equation (8), \( \zeta_1 \) is the location parameter that corresponds to \( p \) and \( g_1(\cdot) \) is the logit link function, while \( \zeta_2 \) is the scale parameter that corresponds to \( \phi \) and \( g_2(\cdot) \) is the logarithm link function. It is straightforward to see that the BBmm and GAMLSS approaches lead to the same definition of the beta-binomial mixed-effects model. However, as will be shown in Section 3.4, there are considerable differences in the estimation process that lead to different parameter estimates.

### 3.4 Differences between BBmm and GAMLSS in the estimation approach

The main difference between the BBmm and GAMLSS methodologies is the estimation procedure. While the GAMLSS methodology is based on an empirical Bayesian argument to derive an estimation procedure, BBmm is based on a classical likelihood framework. It was mentioned in Section 3.3 that GAMLSS models all the parameters of the conditioned distribution by the inclusion of fixed and random effects in the linear predictor. In particular, in the beta-binomial mixed-effects model, both the probability and dispersion parameters are modelled jointly. Nevertheless, due to the model assumptions, the dispersion parameter of the beta-binomial distribution is assumed constant, and hence, it is estimated within the fixed effects estimation procedure. For the GAMLSS model (see Equation (8)) we denote by \( \beta^* = (\beta_1, \beta_2) \) the vector of fixed effects.

Unlike in the BBmm approach, in the GAMLSS methodology the estimation of fixed and random effects is done by means of a posterior mode estimation (Berger, 1985). The model assumes that the joint distribution of all the parameters is given by

\[
f(\beta^*, u, \lambda | y) = f(y | \beta^*, u) f(u | \lambda) f(\lambda) f(\beta^*),
\]

where \( f(y | \beta^*, u) \) is the beta-binomial distribution, \( f(u | \lambda) \) is the normal distribution of the random effects, and \( f(\lambda) \) and \( f(\beta^*) \) are appropriate priors of \( \lambda \) and \( \beta^* \) parameters. Assuming that the hyperparameters or variance components \( \lambda \) are fixed and, assuming a constant improper prior for \( \beta^* \) then the posterior distribution for the fixed and random effects is given by

\[
f(\beta^*, u | y, \lambda) \propto f(y | \beta^*, u) f(u | \lambda).
\]

The estimation of \( \beta^* \) and \( u \) is done by maximising the posterior distribution presented in Equation (9). However, notice that the defined posterior distribution for \( \beta^* \) and \( u \) is exactly the joint log-likelihood in the BBmm approach presented in Equation (4). That is,

\[
\log f(\beta^*, u | y, \lambda) \propto \log f(y | \beta^*, u) + \log f(u | \lambda) = h(\beta, u | y, \theta),
\]

where \( \beta = \beta_1 \) and \( \phi = \exp(\beta_2) \) are included along with \( \lambda \), in the vector of variance components \( \theta \). Consequently, the estimation of the fixed and random effects in GAMLSS is done by maximising the joint log-likelihood exactly as in the BBmm approach. However, there is a crucial difference. In the GAMLSS approach the dispersion parameter of the beta-binomial distribution \( \phi \), or equivalently \( \beta_2 \), is considered as a fixed parameter. Meanwhile, in the BBmm approach it is considered as a variance or dispersion parameter and hence, it is included in the vector of variance parameters \( \theta \). Lee and Nelder (1996) showed that the estimation of dispersion parameters must be done by means of a penalisation of the likelihood in order to avoid bias due to the estimation of the fixed and random effects. The idea was based on the assumption that only the parameters that are canonical can be estimated jointly. Therefore, while in the GAMLSS the estimation of the dispersion parameter of the beta-binomial distribution \( \phi \) is done by maximising the joint log-likelihood, in BBmm the joint log-likelihood is penalised. Indeed, for the estimation of the hyperparameters or variance components of the random effects \( \lambda \), even in the GAMLSS the joint log-likelihood is penalised in order to avoid the bias.
4 | SIMULATION STUDY

In this section, we conduct a simulation study in order to evaluate the performance of the BBmm and GAMlSS approaches when analysing beta-binomial mixed-effects models. The aim of this simulation study is twofold: (a) compare the performance in controlled scenarios given the differences highlighted in Section 3.4 and (b) analyse the estimated parameters in terms of the biases and coverage probabilities of both approaches. The code to implement the BBmm approach has been developed by the authors in the PRDreg package in R, available at CRAN, while the GAMlSS approach is included in the gamlss package in R, version 5.0.1 (Stasinopoulos & Rigby, 2007).

We have generated $R = 100$ random realisations of 200 observations of a dependent variable $y$, which conditional on some simulated random effects follows a beta-binomial distribution with fixed maximum number of scores $m$, the probability parameter $p$, and the dispersion parameter $\phi$. For the sake of clarity, we only consider a single covariate in the linear predictor, which follows a normal distribution of mean 1 and standard deviation 2. Thus, only two fixed effects have been considered, $\beta_0 = 1$ and $\beta_1 = -1.5$. Furthermore, we have considered $q = 50$ realisations of the random effects assuming a normal distribution of mean 0 and standard deviation $\sigma$, where each component is randomly connected from 1 to 9 observations. Note that, according to the general notation in Equation (1), in this case we consider the vector of variance components $\lambda = \sigma$ such that $D = \sigma^2 I_q$. Therefore, the model is defined as

$$\logit(p) = X \beta + Zu,$$

where $\beta = (\beta_0, \beta_1)$, and $u$ is the random effects vector of length $q$.

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_q \end{pmatrix}_{200 \times 2}, \quad \text{where} \quad X_i = \begin{pmatrix} 1 & x_{i,1} \\ \vdots & \vdots \\ 1 & x_{i,n_i} \end{pmatrix}_{n_i \times 2}, \quad \text{and} \quad Z = \begin{pmatrix} 1_{n_1} & 0 & \cdots & 0 \\ 0 & 1_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_{n_q} \end{pmatrix}_{200 \times q}$$

where $1_{n_i}$ is a column vector of 1s of length $n_i$, where $n_i \in \{1, \ldots, 9\}$ and $\sum_{i=1}^{q} n_i = 200$. Indeed, this case study can be considered as a longitudinal study in which each of the 50 individuals has from 1 to 9 repeated realisations of an event.

The simulation study has been divided into several scenarios that depend on the variability of the random effects and the beta-binomial distribution. We consider three possible values, 0.5, 1, and 1.5, for the dispersion parameters $\phi$ and $\sigma$, and hence, a total of nine scenarios with all the possible combinations are defined. For illustrative purposes, Figure 1 shows the shapes of the beta-binomial responses for $x = 1$ and $m = 30$ for each scenario. Additionally, 10 random realisations of the random effects are also shown.

Although the estimates of all the parameters have been obtained by both methodologies, only results for $\beta_1$ and $\phi$ will be shown in detail. First, because the slope shows the relationships between the outcome and individual characteristics, which is the focus of interest in many longitudinal studies. Last, because the estimation of $\phi$ is the main difference between both aforementioned approaches. Nevertheless, the adequacy of the estimates for the other parameters has also been checked.

The results of the simulation study are summarised in Table 2. It is worth mentioning that some convergence problems were reported in the implementation of GAMlSS. We used the default method in the gamlss function, method=RS(), as detailed in Stasinopoulos and Rigby (2007). To compare both approaches, we consider those 100 realisations where both methods converged. Table 2 shows the numerical results based on the following statistics: the mean, the empirical standard deviation (ESD), the average standard deviation (ASD), and the mean squared error (MSE). The coverage probability in % of the 95% Wald confidence intervals for the estimates of $\beta_1$ are also shown in the comparison, since those are the ones reported by gamlss. However, note that from the skewed distributions, the confidence intervals based on the estimates of asymptotic normality can be inaccurate (Royston, 2007). Alternative confidence intervals would be the score or likelihood ratio based intervals (see Agresti, 2007, for further details). The degree of convergence for both methods is also reported in the last column of Table 2.

First, for the slope parameter $\beta_1$, Table 2 shows that the BBmm approach provides much less biased estimates than GAMlSS (given the true value of $\beta_1$ of $-1.5$). In addition, for the BBmm approach, both the ASD and ESD remain quite similar in all the scenarios. In contrast, for GAMlSS, the results are more contradictory (the ASD is always larger than the ESD). Hence, it seems that the GAMlSS approach tends to inflate the standard deviation of the estimates, and consequently, the confidence intervals for $\beta_1$ are larger than they should be, which makes the coverage probability rather high. Furthermore, although the confidence intervals in GAMlSS are overestimated, it can be seen in Table 2 that, as $\sigma$ increases (due to the biased estimation of GAMlSS), the BBmm approach gets better coverage probability results for the slope (especially for $\sigma = 1.5$). Therefore, we can conclude that as the variance of the random effects increases, the results provided by the GAMlSS worsen in terms of
the statistical significance of the parameters. Similar patterns are found for the MSE. While for $\sigma = 0.5$ both methodologies perform similarly, when $\sigma$ increases, there are remarkable differences between both approaches. Results of the simulation study in a graphical display have been included as supplementary material.

Simulations have been performed for larger sample sizes too, namely 500 and 1,000 observations. The results obtained were consistent, showing the same pattern regardless of sample size. Some selected results of the additional simulations are shown as supplementary material.

In summary, based on the results of the simulation study, we can state that although both methodologies perform similarly with the low variance parameter $\sigma$, the results provided by the $BBmm$ approach are better as $\sigma$ increases. Hence, we propose the use of the $BBmm$ methodology as a unified way to analyse real data using a beta-binomial mixed-effects model.

5 | APPLICATION TO COPD STUDY

In this section, we analyse the HRQoL data collected in the COPD study with a beta-binomial mixed-effects regression model based on the $BBmm$ approach. In fact, we apply a longitudinal model where we introduce random intercepts at the baseline and random slopes in the evolution of the patients. In this way, we accommodate the hierarchical structure that exists among the different observations of the same patient. Additionally, we introduce the subtype classification in the model in order to assess the evolution and initial HRQoL of patients clustered in different subtypes based on their health-statuses at the baseline. The model is defined as follows

$$Y_{ij}|u_i, v_i \sim BB(m_{ij}, p_{ij}, \phi)$$

$$\eta_{ij} = \log \frac{p_{ij}}{1 - p_{ij}} = (\beta_0 + \beta_{1k}S_k + u_i) + (\beta_2 + \beta_{3k}S_k + v_i)T_{ij} \quad \text{with } k = 1, 2, 3,$$  

(10)
where \( p_{ij} \) is the probability parameter of the \( i \)th patient in the \( j \)th measurement, \( S_k \) for \( k = 1, 2, 3 \) are the dummy variables for the subtype of the \( i \)th individual, with \( A \) subtype as the reference. \( T_{ij} \) is the day of the measurement of the \( i \)th patient in the \( j \)th time point on the years scale, \( \beta' = (\beta_0, \ldots, \beta_3) \) are the fixed effects, and \( u_i \) and \( v_i \) are, respectively, the random intercept and slope realisation of the \( i \)th individual, \( j = 1, \ldots, n_j, i = 1, \ldots, n \). We assume that \( u_i \) and \( v_i \) have a normal distribution with mean 0 and variance equal to \( \sigma_u^2 \) and \( \sigma_v^2 \), respectively.

Table 3 shows the results of the model defined in Equation (10) for role physical, general health, and role emotional dimensions in the COPD study. Several conclusions can be obtained from the real data application. First, it is worth noting that the differences among the mean health-status of the subtypes at the baseline are statistically significant across the three dimensions. For instance, considering the role emotional dimension, the respective odds ratios for a worse initial emotional state are 1.85 \((1/\exp(-0.618) = 1.85)\), 4.39, and 3.03 for subtypes B, C, and D versus subtype A. Second, with respect to the evolution, Table 3 shows that the patients in subtype A have a statistically significant worsening over time in the role physical and general health dimensions and, furthermore, that there are no statistically significant differences between the evolutions of the patients in different subtypes. Therefore, we can state that, in spite of the initial health-status, patients evolve similarly over time. For example, each year of evolution is associated with a 16.3% worsening in the role physical dimension, whereas it is respectively associated with a 16.8% (i.e., \( 1/\exp(-0.151 - 0.004) = 1.168)\), 15.4%, and 12.3% of worsening for patients in subtypes B, C, and D in which the differences are not statistically significant. Finally, in terms of the variability of the model, Table 3 shows that the largest heterogeneity among patients at the baseline in the temporal evolution is obtained in the role emotional dimension \((\sigma_u = 2.08 \text{ and } \sigma_v = 0.51)\), whereas the general health dimension is the most homogeneous \((\sigma_u = 0.76 \text{ and } \sigma_v = 0.06)\).

Therefore, we can conclude that there are statistically significant differences among patients with COPD at the baseline when divided by subtypes. In fact, patients in subtype A show the best HRQoL in the analysed three dimensions, whereas those in subtype C obtain the worst results in terms of HRQoL. Additionally, we have shown that there are no differences in the evolution among the patients of different subtypes. Hence, we have highlighted that the initial health-status does not determine the evolution of the patients with COPD. Moreover, it is important to note that, whereas in the role emotional patients with COPD do not evolve, the evolution of the role physical and the general health statistically worsen in all the subtypes.
Table 3 Results for the longitudinal beta-binomial model including cluster classification applied to three SF-36 dimensions provided in the COPD study

| Dimension       | Covariate                    | Estimate | SD   | OR   | p-value |
|-----------------|------------------------------|----------|------|------|---------|
| Role physical   | (Intercept)                  | 1.947    | 0.157| –    | <0.001  |
|                 | Year                         | −0.151   | 0.047| 0.860| 0.001   |
|                 | Subtype B                    | −0.942   | 0.198| 0.390| <0.001  |
|                 | Subtype C                    | −2.011   | 0.243| 0.134| <0.001  |
|                 | Subtype D                    | −1.533   | 0.243| 0.216| <0.001  |
|                 | Year × Subtype B             | −0.004   | 0.061| –    | 0.942   |
|                 | Year × Subtype C             | 0.008    | 0.077| –    | 0.921   |
|                 | Year × Subtype D             | 0.035    | 0.083| –    | 0.675   |
|                 | \( \sigma_u \)               | 1.489    | 0.058| –    | –       |
|                 | \( \sigma_v \)               | 0.293    | 0.017| –    | –       |
|                 | \( \log(\phi) \)             | −0.807   | 0.076| –    | –       |
| General health  | (Intercept)                  | 0.114    | 0.043| –    | 0.007   |
|                 | Year                         | −0.050   | 0.013| 0.951| <0.001  |
|                 | Subtype B                    | −0.357   | 0.058| 0.700| <0.001  |
|                 | Subtype C                    | −0.786   | 0.076| 0.456| <0.001  |
|                 | Subtype D                    | −0.591   | 0.075| 0.553| <0.001  |
|                 | Year × Subtype B             | 0.003    | 0.018| –    | 0.861   |
|                 | Year × Subtype C             | 0.044    | 0.025| –    | 0.077   |
|                 | Year × Subtype D             | 0.019    | 0.025| –    | 0.457   |
|                 | \( \sigma_u \)               | 0.759    | 0.024| –    | –       |
|                 | \( \sigma_v \)               | 0.059    | 0.004| –    | –       |
|                 | \( \log(\phi) \)             | −3.780   | 0.123| –    | –       |
| Role emotional  | (Intercept)                  | 2.537    | 0.236| –    | <0.001  |
|                 | Year                         | −0.045   | 0.071| 0.956| 0.528   |
|                 | Subtype B                    | −0.618   | 0.300| 0.539| 0.039   |
|                 | Subtype C                    | −1.478   | 0.351| 0.228| <0.001  |
|                 | Subtype D                    | −1.108   | 0.361| 0.330| 0.002   |
|                 | Year × Subtype B             | −0.116   | 0.093| –    | 0.210   |
|                 | Year × Subtype C             | −0.127   | 0.114| –    | 0.267   |
|                 | Year × Subtype D             | −0.030   | 0.130| –    | 0.815   |
|                 | \( \sigma_u \)               | 2.078    | 0.096| –    | –       |
|                 | \( \sigma_v \)               | 0.513    | 0.032| –    | –       |
|                 | \( \log(\phi) \)             | −0.047   | 0.093| –    | –       |

OR: Odds-ratio; SD: Standard deviation. Subtype A was stated as reference.

Finally, when the GAMLSS approach was applied to data of the COPD study, results were very similar to the aforementioned results obtained with the BBmm approach for the general health dimension. However, for the other two dimensions reported in this work, namely role emotional and role physical, estimation procedure with GAMLSS did not converge.

6 | DISCUSSION

PRO analysis offers the possibility of measuring the effects of some characteristics on a specific disease from the patients’ point of view, which is quite common in many biological or medical studies. Indeed, PROs have gained relevance compared to other
outcomes, such as clinical, physiological, or caregiver-reported outcomes (Deshpande, Rajan, Sudeepthi, & Nazir, 2011). However, the methodologies for analysing this type of data are not common in the literature since most of the regression approaches are based on an assumption that PROs do not satisfy, which is that they do not follow an exponential family distribution. In addition, it is usual to measure PROs over time using a longitudinal study in which the objective lies in the estimation of an outcome evolution for a specific population. This framework further reduces the number of appropriate methodologies in the literature since, apart from dealing with nonexponential family distributions, they must take into account the correlation structure of the repeated measurements.

In this work, we have developed and proposed an estimation procedure named BBmm for the analysis of correlated discrete and bounded outcomes, especially PROs, using a beta-binomial mixed-effects model. Our proposal includes Gaussian random effects to accommodate the correlation structure of a marginal beta-binomial regression model in a longitudinal framework. In addition, we have implemented the methodology in the PRoReg package in R, available at CRAN (R Core Team, 2017).

Although the beta-binomial mixed-effects model that we present is not novel, the estimation procedure that we propose is different from other strategies in the literature, and we show an improvement in the estimates in terms of biases and coverage. In particular, we have compared the BBmm approach to the GAMLSS approach, which is similar to our proposal when restricted to the beta-binomial marginal distribution of the outcome. The main difference between the BBmm and the GAMLSS approaches lies in the penalisation of the joint log-likelihood in order to estimate the dispersion parameter of the beta-binomial distribution. The simulation study shows that in most of the scenarios, the BBmm approach obtained better performance with a smaller bias and a greater coverage probability of the parameters. The improvement is noticeable both in the estimation of the dispersion parameter and in the regression coefficients. Therefore, we conclude that the penalisation approach improves the results in terms of the biases and statistical significances of the estimated regression coefficients. Indeed, the BBmm approach is less likely to experience problems associated with dependent regression parameters in terms of orthogonality, which could result in extra bias in the final results. In terms of the computations, as remarked in the simulation study, we experienced convergence problems in the implementation of GAMLSS using the gamlss R function (particularly for \( \sigma = 1.5 \)). In contrast, our implementation in the BBmm function in the R-package PRoReg always converged. Moreover, the application to real data in the COPD study leads to clinically relevant results.

It is worth mentioning that GAMLSS is a general framework for fitting regression-type models that are conditioned on some given Gaussian random effects. It does not restrict the distribution of the outcomes to exponential family members. Hence, it is not specific for the beta-binomial distribution. Indeed, GAMLSS allows for all the parameters of the distribution of the response variable to be modelled as linear, nonlinear, or smooth functions of the explanatory variables. For the beta-binomial case, it allows for the possibility of specifying a model on the dispersion parameter or in terms of a GAMLSS model \( \eta_2 \) in Equation (8). At this stage, the BBmm approach does not include the possibility of modelling or including smooth effects. However, the estimation procedure developed in this work for the mixed-effects model formulation may include additive nonlinear effects using splines basis functions (Ruppert, Wand, & Carroll, 2003). Moreover, we believe that the strategy that we propose in this work for the penalisation of the joint log-likelihood in the estimation of the dispersion parameters of the conditional distribution, even when they are not canonical to the mean parameters, could be applied in other distributions and it might improve the results in the GAMLSS framework.

As concluding remarks, we recommend the use of the beta-binomial mixed-effects model with the BBmm estimation procedure when analysing longitudinal PROs, provided that the fit to the beta-binomial distribution is adequate, especially when the main objective is to detect factors with significant effects on the evolution of the outcome.

For reproducibility of the methods developed, we provide the R code used in the simulation study and the analysis of real data as supplementary material. Moreover, all R functions, codes and data are available at http://idaejin.github.io/software

ACKNOWLEDGEMENTS

This research was supported by the Basque Government through the BERC 2014–2017 and 2018–2021 programs and the Department of Education, Language Policy and Culture of the Basque Government IT-620-13 programs and Basque Government Industry Department under the ELKARTEK Program, by the Spanish Ministry of Economy and Competitiveness MINECO and FEDER: BCAM Severo Ochoa excellence accreditation SEV-2013-0323, MTM2013-40941-P, MTM2014-52184-P, MTM2016-74931-P, and MTM2017-82379-R funded by (AEI/FEDER, UE) and acronym “AFTERAM”, and by grants from the Instituto de Salud Carlos III, and by the European Regional Development Funds (RD12/0001/0001)—through the thematic networks—REDISSEC (Red de Investigación en Servicios de Salud en Enfermedades Crónicas).
CONFLICT OF INTEREST
The authors have declared no conflict of interest.

ORCID
Dae-Jin Lee https://orcid.org/0000-0002-8995-8535

REFERENCES
Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.
Arostegui, I., Núñez-Antón, V., & Quintana, J. M. (2007). Analysis of the short form-36 SF-36: The beta-binomial distribution approach. *Statistics in Medicine*, 26, 1318–1342.
Arostegui, I., Núñez-Antón, V., & Quintana, J. M. (2012). Statistical approaches to analyse patient-reported outcomes as response variables: An application to health-related quality of life. *Statistical Methods in Medical Research*, 21, 189–214.
Arostegui, I., Núñez-Antón, V., & Quintana, J. M. (2013). On the recoding of continuous and bounded indexes to a binomial form: An application to quality-of-life scores. *Journal of Applied Statistics*, 40, 563–582.
Berger, J. O. (1985). *Statistical decision theory and bayesian analysis*. New York: Springer.
Bernardo, J., & Smith, A. (1994). *Bayesian theory* (Chapter 5, pp. 240–376). Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.
Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25.
Buist, A. S., Vollmer, W. M., & McBurnie, M. A. (2008). Worldwide burden of COPD in high-and low-income countries. Part I. The burden of obstructive lung disease (BOLD) initiative. *International Journal of Tuberculosis and Lung Disease*, 12, 703–708.
Deshpande, P. R., Rajan, S., Sudeepthi, B. L., & Nazir, C. P. A. (2011). Patient-reported outcomes: A new era in clinical research. *Perspectives in Clinical Research*, 2, 137–144.
Dickey, J. (1982). In *Encyclopedia of statistical sciences*, (Vol. 1, pp. 135–145.). Hoboken, New Jersey: John Wiley & Sons, Inc.
Esteban, C., Arostegui, I., Aburto, M., Moraza, J., Quintana, J. M., García-Loizaga, A., … Capelastegui, A. (2016). Chronic obstructive pulmonary disease subtypes. Transitions over time. *PLoS One*, 11(9), e0161710.
Garratt, A. M., Schmidt, L., Mackintosh, A., & Fitzpatrick, R. (2002). Quality of life measurement: Bibliographic study of patient assessed health outcome measures. *British Medical Journal*, 324, 1417–1421.
George, K., Iddi, S., & Molenberghs, G. (2016). The combined model: A tool for simulating correlated counts with overdispersion. *Communication in Statistics- Simulation and Computation*, 45, 2491–2510.
McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall/CRC.
McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.
McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear and mixed models*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
Molenberghs, G., Verbeke, G., Demétrio, C., & Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, 25(3), 325–347.
Molenberghs, G., Verbeke, G., Iddi, S., & Demétrio, C. (2012). A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis*, 111, 94–109.
Murray, C. J., & López, A. D. (1997). Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. *Lancet*, 349, 1498–504.
Najera-Zuloaga, J., Lee, D.-J., & Arostegui, I. (2017). Comparison of beta-binomial regression model approaches to analyze health-related quality of life data. *Statistical Methods in Medical Research*, 27(10), 2989–3009.
Patterson, H. D., & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545–554.
SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Najera-Zuloaga J, Lee DJ, Arostegui I. A beta-binomial mixed-effects model approach for analysing longitudinal discrete and bounded outcomes. Biometrical Journal. 2019;61:600–615. https://doi.org/10.1002/bimj.201700251

APPENDIX: THE BETA-BINOMIAL DISTRIBUTION

The beta-binomial distribution consists of a finite sum of correlated binary outcomes whose probability parameter is assumed to be random and drawn from a beta distribution. Let $y_1, \ldots, y_m$ be a set of binary outcomes and $\psi$ a random variable following a beta distribution with parameters $\alpha_1, \alpha_2 > 0$. Additionally, assume that conditional on $\psi$, the binary outcomes are independent and identically distributed (iid) following a Bernoulli distribution with the probability parameter $\psi$. Namely,

$$y_j | \psi \sim \text{Ber}(\psi) \quad \text{iid, where} \quad \psi \sim \text{Beta}(\alpha_1, \alpha_2), \quad j = 1, \ldots, m,$$

with $E(\psi) = p$ and $\text{Var}(\psi) = p(1 - p) \phi / (1 + \phi)$, where $p = \alpha_1 / (\alpha_1 + \alpha_2)$ and $\phi = 1 / (\alpha_1 + \alpha_2)$. Hence, the marginal first- and second-order moments of the outcome variables are defined as

$$E(y_j) = E\left[E(y_j | \psi)\right] = p$$
$$\text{Var}(y_j) = \text{Var}\left[E(y_j | \psi)\right] + E\left[\text{Var}(y_j | \psi)\right] = p(1 - p), \quad j = 1, \ldots, m.$$ 

It can be noticed that marginal moments correspond to the usual Bernoulli distribution moments. However, the assumption of a random probability parameter connects the observations through the intraclass correlation parameter $\rho$, which is defined as

$$\rho = \text{Corr}(y_j, y_k) = \frac{\text{Cov}(y_j, y_k)}{\sqrt{\text{Var}(y_j)}\sqrt{\text{Var}(y_k)}} = \frac{\phi}{1 + \phi},$$

where $k, j = 1, \ldots, m$ and $j \neq k$. This correlation formula determines the parameter $\phi$ as a dispersion or correlation parameter.
If we sum up all the correlated binary outcomes, we will define a new variable as

$$y = \sum_{j=1}^{m} y_j,$$

which follows the beta-binomial distribution. Indeed, a random variable $y$ follows a beta-binomial distribution with the parameters $m$, $p$, and $\phi$ if

$$y|\psi \sim \text{Bin}(m, \psi) \quad \text{and} \quad \psi \sim \text{Beta}(p/\phi, (1-p)/\phi).$$

The probability mass of the beta-binomial distribution is given by

$$f(y) = \int_{0}^{1} f_{y|\psi}(y|\psi)f_{\psi}(\psi)d\psi
= \left( \frac{m}{y} \right) \frac{\Gamma(1/\phi)}{\Gamma(1/\phi + m)} \frac{\Gamma(p/\phi + y)}{\Gamma(p/\phi)} \frac{\Gamma((1-p)/\phi + m - y)}{\Gamma((1-p)/\phi)}.$$

(A.1)

Additionally, the first- and second-order moments of the beta-binomial distribution are defined as

$$\text{E}(y) = mp \quad \text{and} \quad \text{Var}(y) = mp(1-p) \left[ 1 + (m-1) \frac{\phi}{1+\phi} \right].$$

Therefore, one can interpret the beta-binomial model as a binomial distribution with an additional variability coming from the existing correlation among the $y_j$'s. More details about the beta-binomial distribution can be found in Johnson, Kemp, and Kotz (2005).