PLOS ONE

RESEARCH ARTICLE

A mixed-effects two-part model for twin-data and an application on identifying important factors associated with extremely preterm children's health disorders

Baiming Zou[1]*, **Hudson P. Santos**[2], **James G. Xenakis**[3], **Mike M. O'Shea**[4], **Rebecca C. Fry**[5], **Fei Zou**[1]

1 Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States of America, **2** School of Nursing, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States of America, **3** Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States of America, **4** Division of Neonatal-Perinatal Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States of America, **5** Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States of America

* bzou@email.unc.edu

Citation: Zou B, Santos HP, Xenakis JG, O'Shea MM, Fry RC, Zou F (2022) A mixed-effects two-part model for twin-data and an application on identifying important factors associated with extremely preterm children's health disorders. PLoS ONE 17(6): e0269630. https://doi.org/10.1371/journal.pone.0269630

Editor: Pedro Vieira da Silva Magalhaes, Universidade Federal do Rio Grande do Sul, BRAZIL

Received: October 11, 2021

Accepted: May 24, 2022

Published: June 13, 2022

Data Availability Statement: All relevant data are within the paper and its Supporting information files.

Funding: This study was supported by grants from the National Institutes of Health (NIH), specifically the National Institute of Neurological Disorders and Stroke (U01NS040069; R01NS040069), the Office of the NIH Director (UG3OD023348), the National Institute of Environmental Health Sciences

Abstract

Our recent studies identifying factors significantly associated with the positive child health index (PCHI) in a mixed cohort of preterm-born singletons, twins, and triplets posed some analytic and modeling challenges. The PCHI transforms the total number of health disorders experienced (of the eleven ascertained) to a scale from 0 to 100%. While some of the children had none of the eleven health disorders (i.e., PCHI = 1), others experienced a subset or all (i.e., $0 \leq$PCHI< 1). This indicates the existence of two distinct data processes—one for the healthy children, and another for those with at least one health disorder, necessitating a two-part model to accommodate both. Further, the scores for twins and triplets are potentially correlated since these children share similar genetics and early environments. The existing approach for analyzing PCHI data dichotomizes the data (i.e., number of health disorders) and uses a mixed-effects logistic or multiple logistic regression to model the binary feature of the PCHI (1 vs. < 1). To provide an alternate analytic framework, in this study we jointly model the two data processes under a mixed-effects two-part model framework that accounts for the sample correlations between and within the two data processes. The proposed method increases power to detect factors associated with disorders. Extensive numerical studies demonstrate that the proposed joint-test procedure consistently outperforms the existing method when the type I error is controlled at the same level. Our numerical studies also show that the proposed method is robust to model misspecifications and it is applicable to a set of correlated semi-continuous data.

(T32ES007018; T32ES007126; P42ES031007), National Institute of Nursing Research (K23NR017898; R01NR019245), and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01HD092374; R03HD101413).

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Approximately 10% of babies are born prematurely (< 37 weeks gestational age) worldwide every year [1]. Preterm birth is the leading cause of neonatal mortality and an important risk factor for developmental impairments including cognitive, behavioral, and social-emotional disorders [2–7]. Individuals born preterm also experience a higher risk of health disorders (e.g., asthma) and have a shorter life expectancy [8]. Our team has investigated 11 health disorders that are associated with preterm birth including bilateral blindness, hearing impairment, moderate/severe cognitive impairment, epilepsy, gross motor function impairment, attention-deficit/hyperactivity disorder, anxiety, depression, asthma, autism, and obesity (i.e., body mass index above the 95 percentile) [9]. Further, preterm birth in the U.S. is associated with billions of dollars of annual societal economic burden [10]. On the other hand, some preterm children do not display major adverse health or developmental outcomes [9, 11]. Investigating the outcomes of preterm-born children remains critical for evaluating and improving clinical care, planning long-term support and for advancing our understanding of the life-course consequences of immaturity at birth [7]. Identifying important factors that are associated with the risk and burden of such health disorders is important to both clinicians and researchers [12]. Indeed, determining important biomarkers associated with adverse health outcomes has long been recognized as a critical component in investigating disease etiologies, developing new therapeutic interventions, and accurately predicting disease progression [13, 14]. However, the identification of such biomarkers remains a challenge.

The current research is motivated by recent epidemiological studies investigating the positive child health outcomes among 10-year-old children who were born extremely preterm [9, 12]. One of our primary scientific interests in these studies was to identify important factors associated with the positive child health index (PCHI) outcome, which summates information about the presence or absence of 11 adverse health disorders at age 10 and transforms the cumulative number to a scale from 0 (the child experienced all the disorders) to 100% (no disorders), accounting for the number of non-missing responses [12]. That is, the fewer disorders the child experienced, the higher the child's PCHI score. Two aspects of the PCHI lead to analytic and modeling challenges when investigating associations between maternal antecedents, child's characteristics and PCHIs. The first involves the semi-continuous nature of the scale, which clearly reveals two distinct underlying data processes. While some children experienced none of these disorders (i.e., PCHI = 1), many others experienced a subset or all (i.e., PCHI < 1). That is, the PCHI, or alternatively the number of health disorder measurements, follows a mixture distribution, taking a boundary value under one condition, or a value arising from a continuous (ordinal) distribution under another. The second challenging feature of these data was that, in addition to singletons, it included twin and triplet clusters, within which PCHI measurements, or equivalently the number of health disorder experienced by each of the twins and triplets were correlated, as members of the same nuclear family share similar genetic structures and are exposed to comparable early environments. Even more challenging is the fact that these family-related cluster effects which within the two underlying data processes impact the propensity of incurring adverse health outcome and their subsequent burden, respectively, are themselves usually correlated with each other.

Previous research employed a strategy to identify important risk factors associated with the PCHI by dichotomizing the scale based on the number of disorders experienced (no any health disorder versus at least one health disorder, i.e., PCHI= 1 versus PCHI< 1) and using this binary outcome in a mixed-effects logistic regression with a random intercept to correlate children from a multiple birth [12]. Although this modeling scheme accommodates all the samples and is mathematically convenient, it can limit statistical power, since it is only capable of

capturing the impact of risk factors on a binary health disorder status. However, in practice, these risk factors might also impact the disease burden, i.e., these risk factors may not only influence the propensity of developing health disorders but also impact the burden of disease (i.e., the number of health disorder experienced). More importantly, these risk factors usually impact both positively or negatively on the two data processes, i.e. in the same direction. Thus, jointly modeling the effects of these factors on the health outcomes can boost detection power. For independent samples, to jointly model this type of semi-continuous data, a two-part model can be adopted [15–20], where a logistic regression model is used to model the binary event of incurring no health disorders versus incurring at least one disorder, while a multiple linear (log-normal) regression is employed to model the association between the factors and the burden of health disorders (i.e., the nonzero data where the number of disorder experienced is converted to the proportion of the total 11 disorder, the larger the more severe). Special treatment is still required to account for the correlations among twin or triplet samples. To this end, we adopt the mixed-effects two-part modeling framework developed for longitudinal semi-continuous data [21]. Jointly, we model the correlated semi-continuous data by taking into account the sample correlations between and within the two underlying data processes. Under this framework, we derive a joint-test procedure for assessing each risk factor's effect on the correlated semi-continuous outcome measurements.

The remainder of the paper is arranged as follows. In Section 2, we provide a detailed description of the proposed mixed-effects two-part modeling framework for the correlated semi-continuous data. We then apply the proposed joint-test procedure to identify important factors associated with the health disorders for extremely preterm children in Section 3 to demonstrate its benefits over the existing method. We conduct extensive numerical studies under different data complexity scenarios in Section 4 to explore the applicability and performance as compared with the existing method and to demonstrate the generalizability and robustness of the proposed joint-test procedure for the correlated semi-continuous data. The paper concludes with some discussions in Section 5.

2 Joint-test procedure under mixed-effects two-part model for correlated semi-continuous data

Before providing more detailed explication of our proposed joint-test procedure, we first introduce some notations. Let $\mathbf{X}$ represent a matrix of all the observed clinical, demographic and biomarker variables, and Y represent an outcome vector of interest (e.g., the health disorder score), which can be zero or a continuous positive value. In the context of PCHI, equivalently, $Y = 1 - \text{PCHI}$ where zero corresponds to experiencing no any adverse health issues while a positive value represents experiencing the proportion of 11 adverse health disorders (the larger the value, the more adverse outcomes experienced). We introduce another binary variable vector, Z, which is a dichotomization of the adverse health status, such that $Z = \mathbf{I}(Y > 0)$ where $\mathbf{I}$ is an indicator function. Our primary scientific interest is to model $\Pr(Y > 0 \mid \mathbf{X})$ and $\mathrm{E}[Y|Y > 0, \mathbf{X})]$ such that we can identify risk factors that are significantly associated with the propensity of being in disease status (i.e., $Z = 1$) and the burden of disease (i.e., $Y > 0$) when in disease state.

To model the association between the observed risk factors $\mathbf{X}$ and the adverse health status Z for the correlated data, we adopt a mixed-effects multiple logistic regression model, while the impacts of risk factors on the burden of the adverse health outcome are modelled with a log-normal distribution as follows, although other mixed-effects (e.g., linear or Poission)

models can be employed. Our models can be written as follows:

$$\text{logit}(\pi_{ij}) \quad = \quad \alpha_0 + \sum_{k=1}^{p} \alpha_k x_{k,ij} + u_i \tag{1}$$

$$\log(y_{ij})|y_{ij} > 0 \quad = \quad \beta_0 + \sum_{k=1}^{p} \beta_k x_{k,ij} + v_i + \epsilon_{ij} \tag{2}$$

where $x_{k,ij}$ is the k^{th} observed risk factor of the j^{th} subject in family i ($i = 1, \cdots, m$; $j = 1, \cdots, n_i$), $\pi_{ij} \equiv \Pr(z_{ij} = 1|\{\mathbf{x}_{ij}, u_i\})$ and $\epsilon_{ij} \sim N(0, \sigma^2)$ represents a random error term. The family-specific effects on the propensity to have at least one of the health disorders, and the proportion of having all health disorders (conditional on having at least one of them) are represented by u_i and v_i, respectively; to characterize the correlation between them, we assume they follow a bivariate normal distribution with mean zero as following:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix} \right).$$

We can write down the total likelihood function as follows:

$$\begin{aligned} L \quad &= \quad \prod_{i=1}^{m} \int_{u_i} \int_{v_i} \prod_{j=1}^{m_i} f(z_{ij}, y_{ij}|\boldsymbol{\alpha}, \boldsymbol{\beta}, u_i, v_i, \sigma^2) f(u_i, v_i|\sigma_u^2, \sigma_v^2, \rho) dv_i du_i \\ &= \quad \prod_{i=1}^{m} \int_{u_i} \int_{v_i} \prod_{j=1}^{m_i} \{1 - Pr(z_{ij} = 1|\boldsymbol{\alpha}, u_i)\}^{(1-z_{ij})} \{Pr(z_{ij} = 1|\boldsymbol{\alpha}, u_i)\}^{z_{ij}} \\ &\quad [f(y_{ij}|\boldsymbol{\beta}, v_i, \sigma^2)]^{z_{ij}} f(u_i, v_i|\sigma_u^2, \sigma_v^2, \rho) dv_i du_i \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_p)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$, representing the risk factor effects on the propensity of developing at least one health disorder and the effects on the disease burden (e.g., characterized by the proportion of 11 health disorders involved) conditional on the subject developing the health disorder, respectively, and $f(\cdot)$ denotes a density function. Estimates for the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \sigma_u^2, \sigma_v^2, \rho)$ can be obtained by maximizing the above likelihood function. However, this optimization could be challenging since the likelihood function involves two intractable integrals, which usually cannot be dealt with separately, except in the case of independent data [22]. Several strategies can be adopted, for example, the Fisher score approach via Laplace approximation [21], or the quasi-Newton optimization via adaptive Gaussian quadrature [23, 24]. In this paper, we adopt a hybrid expectation maximization (EM) and quasi-Newton algorithm [25].

To test the k^{th} hypothesis: $H_k^0 : \alpha_k = \beta_k = 0$ vs H_k^1 : otherwise $(k = 1, \cdots, p)$, we construct the following test statistic:

$$\zeta = (\hat{\alpha}_k, \hat{\beta}_k)\Sigma^{-1} \begin{pmatrix} \hat{\alpha}_k \\ \hat{\beta}_k \end{pmatrix} \tag{4}$$

where $\hat{\alpha}_k$ and $\hat{\beta}_k$ are the maximum likelihood estimates (MLEs) for parameters α_k and β_k, respectively, from Model (3). They follow a bivariate normal distribution with mean zero

under the null as follows:

$$\begin{pmatrix} \hat{\alpha}_k \\ \hat{\beta}_k \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right) \tag{5}$$

where $\Sigma = \begin{pmatrix} \sigma^2_{\alpha_k} & \varrho\sigma_{\alpha_k}\sigma_{\beta_k} \\ \varrho\sigma_{\alpha_k}\sigma_{\beta_k} & \sigma^2_{\beta_k} \end{pmatrix}$ is the covariance matrix between $\hat{\alpha}_k$ and $\hat{\beta}_k$, with $\sigma^2_{\alpha_k}$ and $\sigma^2_{\beta_k}$ being the variance of the MLEs, respectively, and ϱ being the correlation between these MLEs. Σ can be estimated by maximizing the likelihood function (3) as well. Test statistic ζ follows a $\chi^2(2)$ distribution under the null.

In summary, the existing analysis method for this type of correlated semi-continuous PCHI data, i.e. Model (1), only models the relationship between the risk factors and the chance to be in positive health disorder status by dichotomizing PCHI scores (i.e., PCHI = 1 vs PCHI$<$ 1). However, the risk factors usually not only influence the chance of being in health disorder status but also impact the disease burden. The proposed analysis method, i.e. Model (3), considers the risk factors effect on both the probability and burden of being in health disorder by incorporating Models (1) and (2). Furthermore, although this is a cross-sectional study (i.e. PCHI scores measured at age of 10-year), the proposed analysis method adopted mixed-effect models for longitudinal data to take into account the correlations between and within the twin's and triplet's PCHI (health disorder) measurements in the two data processes.

3 A practical application

To investigate the performance of the proposed method in practice, we apply it to the analysis of data from the Extremely Low Gestational Age Newborn (ELGAN) study, which has motivated this research. The ELGAN study is a prospective longitudinal observational cohort study. Study participants were enrolled at 14 hospitals in 5 states in the United States (Connecticut, Massachusetts, Illinois, Michigan, and North Carolina) between April 2002 and August 2004. The only inclusion criteria were birth at one of the enrollment hospitals and birth before 28 completed weeks of gestation [26, 27]. The only exclusion criterion was anencephaly. Longitudinal follow-up of the 1198 surviving participants in the ELGAN cohort has included research clinic visits at 1, 2, 10, and 15 years of age. Children were diagnosed with cerebral palsy based on a standardized assessment at 2 years of age, adjusted for degree of prematurity [28]. All other assessments that were used to derive the positive child health index occurred when study participants were 10 years of age. Parent report was used to identify bilateral blindness and hearing impairment. Children's weight and height were measured by research coordinators at 10 years of age and body mass index (BMI) percentiles were calculated based on measured weight and height and age- and sex-specific US growth standards; obesity was defined as a BMI percentile $\geq$ 95% [29]. Asthma diagnosis at age 10 years was based on parent or guardian report of a health care provider's diagnosis of asthma [30]. Here, we focus on estimating the impacts of sex, race, birth weight, maternal pre-pregnancy BMI, maternal smoking status during pregnancy, maternal age, maternal education level, histologic chorioamnionitis (a prenatal infection of the fetal membranes), pre-pregnancy maternal asthma status, and insurance status (public health insurance at birth) on the health disorders measured at age of 10 year-old for extremely preterm-born children, and identify the important factors associated with the health disorders. The data consist of 776 complete samples in total, of which 550 are singletons, 188 are twins (94 pairs), 33 are triplets (11 triplets), and 5 are quintuplets. The distributions of all observed clinical and demographic variables are

Table 1. Distribution of child and maternal characteristics.

Child Characteristics		N or Mean	Percentage or SD
Number of disorder	None	251	32.3
	At least one	525	67.7
Sex	Female	369	47.6
	Male	407	52.4
Race	White	498	64.2
	Other	278	35.8
Birth weight		835.8	197.6
Maternal Characteristics		**N or Mean**	**Percentage or SD**
Smoking status during pregnancy	Yes	107	13.8
	No	669	86.2
Pre-pregnancy asthma	Yes	92	11.9
	No	684	88.1
Public insurance at birth	Yes	269	34.7
	No	507	65.3
Maternal education	$\leq$ 12 years	103	13.3
	$>$ 12 years	673	86.7
Histologic chorioamnionitis	Yes	268	34.5
	No	508	65.5
Pre-pregnancy BMI		25.5	6.9
Maternal age		29.3	6.7

https://doi.org/10.1371/journal.pone.0269630.t001

summarized in Table 1. We note that nearly one third of the subjects (n = 251) reported no any health disorders at age 10.

We compare the performance of our proposed analysis method, i.e. the joint mixed-effects model (3) which jointly tests the effect of each feature using the test statistic (4) as described above, with the existing analysis method for PCHI data, i.e. mixed-effects logistic regression approach (1). In our analysis, pre-pregnancy asthma and histologic chorioamnionitis were included in the two-part model since they are potential confounding variables for the health disorders [12]. The parameter estimates and associated 95% confidence interval (CI) are presented in Table 2. As shown, at the 0.05 level of significance, the conventional approach detects sex, race, maternal pre-pregnancy BMI and public health insurance (rather than private insurance) as important risk factors, while the joint-test procedure identifies sex, birth weight, maternal pre-pregnancy BMI and public health insurance. However, after adjusting for multiple comparisons using the Hommel [31] or Bonferroni correction, only maternal pre-pregnancy BMI remains significant in the conventional method with an adjusted p-value of 0.008 (the adjusted p-values for sex, race and public health insurance become 0.177, 0.189, 0.115, respectively). Conversely, sex, maternal pre-pregnancy BMI and public health insurance remain significant in the joint-test procedure after this adjustment, with p-values of 0.010, 0.005, and 0.005, respectively. Although both methods detect pre-pregnancy BMI as a significant risk factor, we note that the adjusted p-value from the joint-test procedure is much smaller than that from the conventional method (0.005 vs 0.008). The consistency of corresponding pairs of parameter estimates from the mixed-effects two-part model is reassuring. That is, the estimates for sex (0.386;0.143), maternal pre-pregnancy BMI (0.346;0.053), and public health insurance (0.594;0.188), indicate that each influences the propensities of incurring health disorders and the health disorder burden (conditional on at least one health

Table 2. PCHI data analysis results.

	Existing Method		Model (2)		Proposed Method	
Variable	Estimate (α)	95% CI	Estimate (β)	95% CI	Estimate (α;β)	p-value
Sex	0.380	(0.047,0.713)	0.141	(0.049, 0.234)*	0.386; 0.143	0.001*
Race	0.451	(0.051,0.850)	0.036	(-0.068,0.140)	0.498; 0.040	0.067
Birth weight	-0.143	(-0.309,0.023)	-0.050	(-0.096,-0.005)	-0.157;-0.052	0.023
Pre-pregnancy BMI	0.330	(0.137,0.523)*	0.050	(0.005,0.094)	0.346; 0.053	0.001*
Smoking status during pregnancy	0.403	(-0.149,0.955)	0.076	(-0.057,0.209)	0.438; 0.080	0.188
Maternal age	-0.177	(-0.375,0.021)	0.021	(-0.034,0.077)	-0.183; 0.021	0.181
Maternal education	-0.178	(-0.785,0.430)	-0.119	(-0.258,0.019)	-0.173;-0.120	0.209
Histologic chorioamnionitis	0.109	(-0.244,0.462)	0.047	(-0.051,0.144)	0.133; 0.049	0.495
Pre-pregnancy asthma	0.373	(-0.200,0.946)	0.101	(-0.035,0.236)	0.403; 0.102	0.159
Public health insurance at birth	0.564	(0.104,1.025)	0.182	(0.068,0.296)*	0.594; 0.188	0.001*

*statistically significant at significance level 0.05 after adjusting for multiple comparison

https://doi.org/10.1371/journal.pone.0269630.t002

disorder) in the same direction. Even though the parameter estimate (i.e., α) difference between existing method and the proposed method is small, joint-modeling includes the information from the second model (i.e. β). For example, using existing method, birth weight effect estimate is −0.143, which is not statistically significant. However, using the proposed joint-modeling method, the birth weight effect on the propensity of developing health disorder is estimated as −0.157, yet the birth weight is detected as a significant factor (p-value = 0.023) since the joint-modeling method considers the birth weight effect on the disease burden with a parameter estimate of −0.052. This provides intuition to understand why the joint modeling can potentially boost detection power compared with the conventional method. Also, in our analysis, treating pre-pregnancy BMI as a continuous or categorical variable won't change the overall conclusion.

4 Simulation study

The semi-continuous nature of the PCHI data is actually a relatively common feature of scientific and clinical data. For example, medical expenditures and length of hospital stay are two other examples that give rise to such data; in a given year some people accrue no medical expenses (or require no time in hospital) at all. Otherwise, these outcomes are represented by positive numbers (i.e., dollars or days). Examples of such data abound in economic studies—for example, the size of an insurance claim will be a positive number when an incident occurs, and zero otherwise. Other examples include postoperative pain (POP) measured sometime post surgery; many surgery patients experience POP with varying intensity (i.e., the POP scores take positive values), while POP will completely resolve by the measurement occasion for many others (i.e., the POP scores are zero). An example with particular relevance to studies of extremely preterm babies is the duration of mechanical ventilation, which is 0 for a substantial proportion of infants and continuous positive values for others.

We conducted extensive simulation studies to explore the applicability of the joint-test procedure under scenarios of varying data complexity. In our first scenario, we simulated the correlated data using a mixed-effects logistic regression model to generate the binary adverse health status z, and a log-normal linear mixed-effects model to generate the positive data

process, i.e., the burden of the adverse health outcome y, as presented below:

$$\text{logit}(\pi_{ij}) = 0.8 + 0.2x_{1,ij} + 0.4w_{1,ij} + 0.45x_{2,ij} + 0.6w_{2,ij} + u_i$$

$$\log(y_{ij}) = -0.5 + 0.25x_{1,ij} + 0.5w_{1,ij} + 0.4x_{3,ij} + 0.6w_{3,ij} + v_i + \epsilon_{ij} \text{ if } z_{ij} = 1$$

where $\pi_{ij} \equiv \Pr(z_{ij} = 1 | \{x_{1,ij}, x_{2,ij}, w_{1,ij}, w_{2,ij}, u_i\})$ and $\epsilon_{ij} \sim N(0, 1)$ is a random error term. In addition to the causal risk factors (i.e. x_1, x_2, x_3 and w_1, w_2, w_3, which correspond to continuous and binary variables, respectively), we also generate two nuisance factors, one continuous x_4 and the other binary w_4. The continuous variables follow a multivariate normal distribution and the binary variables correlate with the continuous variables $x_4, \cdots, x_8$ as follows:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_8 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 & \cdots & 0.3 \\ 0.3 & 1 & 0.3 & \cdots \\ \vdots & 0.3 & 1 & 0.3 \\ 0.3 & \cdots & 0.3 & 1 \end{pmatrix} \right)$$

$$w_p = \begin{cases} 1 & \text{if } x_{p+4} > 0 \\ 0 & \text{if } x_{p+4} \leq 0 \end{cases} \qquad (p = 1, \cdots, 4)$$

The random-effects terms u_i and v_i follow a bivariate normal distribution as follows:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Using sample sizes of 500 (400 singletons and 50 twin pairs) and 1000 (800 singletons and 100 twin pairs), we report the percentage of 1000 Monte Carlo replications run that identify each of the eight risk factors ($X_1, \cdots, X_4, W_1, \cdots, W_4$) as statistically significant under different correlation settings for ($\rho = 0$ and $\rho = 0.4$) in Tables 3 and 4, respectively, controlling the type I error rate at the 0.05 level.

The results in Tables 3 and 4 demonstrate that the joint-test procedure (i.e., the column labeled "Proposed Method") outperforms the traditional mixed-effects logistic regression method ("Existing Method") in detecting the important factors. For example, when $n = 500$ and $\rho = 0$, the power to detect the risk factors X_1 and W_1 (which influence both the binary events and the continuous positive outcomes) are markedly boosted from 0.298 and 0.332 to 0.859 and 0.880, respectively. Similar observations pertain to the other settings for varying n and ρ. Furthermore, when the traditional method fails to detect risk factors that only influence the continuous outcome, the joint-test procedure still can detect these factors with very high power. For instance, under the setting of $n = 500$ and $\rho = 0.4$, the traditional method fails to detect X_3 and W_3 (the negligible powers 0.059 and 0.059 are essentially equal to the nominal type I error rate), but the joint-test procedure has nearly full power to detect these two risk factors. In addition, the empirical type I error is appropriately controlled.

To further investigate the robustness of our two-part modeling approach to identify important biomarkers for correlated data, we conducted additional simulations using the following data generating models:

$$\text{logit}(\pi_{ij}) = 0.8 + 0.25x_{1,ij} + 0.2w_{1,ij} + 0.4x_{2,ij} + 0.6w_{2,ij} + u_i$$

$$y_{ij} \sim \text{Poisson}(16 + 0.65x_{1,ij} + 0.85w_{1,ij} + 0.5x_{3,ij} + w_{3,ij} + v_i) \text{ if } z_{ij} = 1$$

Table 3. Power and type I error comparison for cluster log-normal data ($\rho = 0$).

	$n = 50$			$n = 1000$		
Variable	Existing Method	Model (2)	Proposed Method	Existing Method	Model (2)	Proposed Method
X_1	0.298	0.400	0.859	0.509	0.597	0.993
W_1	0.332	0.316	0.880	0.562	0.523	1.000
X_2	0.880	0.053	0.729	0.995	0.046	0.988
W_2	0.602	0.035	0.408	0.906	0.038	0.811
X_3	0.051	0.754	0.993	0.055	0.926	1.000
W_3	0.055	0.454	0.956	0.049	0.730	1.000
X_4	0.056	0.059	0.048	0.044	0.052	0.037
W_4	0.040	0.045	0.042	0.056	0.041	0.050

https://doi.org/10.1371/journal.pone.0269630.t003

Table 4. Power and type I error comparison for cluster log-normal data ($\rho = 0.4$).

	$n = 50$			$n = 1000$		
Variable	Existing Method	Model (2)	Proposed Method	Existing Method	Model (2)	Proposed Method
X_1	0.302	0.394	0.841	0.505	0.592	0.991
W_1	0.319	0.302	0.867	0.569	0.515	0.993
X_2	0.893	0.055	0.739	0.994	0.057	0.989
W_2	0.609	0.044	0.414	0.903	0.037	0.813
X_3	0.059	0.783	0.996	0.049	0.933	1.000
W_3	0.059	0.452	0.940	0.053	0.740	1.000
X_4	0.053	0.052	0.034	0.063	0.041	0.043
W_4	0.055	0.039	0.045	0.046	0.036	0.046

https://doi.org/10.1371/journal.pone.0269630.t004

where the causal risk factors, nuisance variables, and cluster effects settings are the same as in the above cluster log-normal scenario. Here, we still employ a multiple linear log-normal mixed-effects model for the positive portion of the data, but these data are now generated using a Poisson mixed-effects model. We ran 1000 Monte Carlo replications with a sample size $n = 1000$ and two different levels of correlation between the cluster effects ($\rho = 0$ and $\rho = 0.4$). The percentage of replications deeming the eight risk factors as statistically significant (while controlling the type I error rate at the 0.05) level are presented in Table 5. Again, the results in

Table 5. Power and type I error comparison for cluster poisson data.

	$\rho = 0$			$\rho = 0.4$		
Variable	Existing Method	Model (2)	Proposed Method	Existing Method	Model (2)	Proposed Method
X_1	0.743	0.945	0.982	0.736	0.927	0.976
W_1	0.192	0.688	0.654	0.183	0.703	0.656
X_2	0.982	0.054	0.953	0.977	0.052	0.950
W_2	0.884	0.047	0.810	0.893	0.050	0.805
X_3	0.056	0.785	0.688	0.054	0.775	0.695
W_3	0.048	0.861	0.756	0.057	0.845	0.752
X_4	0.058	0.052	0.037	0.056	0.047	0.050
W_4	0.043	0.057	0.045	0.053	0.047	0.045

https://doi.org/10.1371/journal.pone.0269630.t005

Table 5 demonstrate that the joint-test procedure under two-part modeling ("Proposed Method") is much more powerful than the traditional one-part mixed-effects logistic regression model ("Existing Method") to detect risk factors while controlling the type I error rate at a reasonable level regardless of the correlation settings. Of note, even when the intensity model (i.e., the log-normal mixed-effects model) is misspecified for the positive ordinal data in the two-part modeling framework, the joint-test procedure still outperforms the traditional approach, further demonstrating its utility in clinical practice.

5 Discussion

In this study, we derived a joint-test procedure under a mixed-effects two-part modeling framework to identify important risk factors associated with the correlated semi-continuous outcomes. The application of the proposed method to the real PCHI data analysis has clearly demonstrated the advantages of the proposed method compared to the existing method for this type of semi-continuous data, i.e., it is more powerful to detect risk factors associated with the correlated semi-continuous outcomes. Further, this method is robust to model misspecification. In general, our proposed joint-test procedure is consistently more powerful than the traditional method while controlling the type I error rate at the same targeted level. The advantages of the proposed two-part model rooted from the fact that it jointly models the two data processes unlike the existing method (e.g., the censored or truncated regression) assumes one underlying data process for all subjects with a ceiling effect and thus is less flexible. These results manifestly demonstrate the advantages of the mixed-effects two-part model for the analysis of correlated semi-continuous data.

In pediatric practice, in addition to identifying important risk factors associated with cross-sectional adverse health outcomes for preterm-born children, it is of clinical importance to investigate their associations with longitudinal health outcomes as well. It is of further practical importance to identify genetic variants that are associated with the health outcomes of these children. How to extend the proposed joint-test procedure to these even more complicated longitudinal correlated semi-continuous (and, in the latter case, high dimensional) settings warrants further investigations, but is beyond the scope of this paper.

Supporting information

S1 Data.
(CSV)

Author Contributions

Data curation: Hudson P. Santos, Mike M. O'Shea, Rebecca C. Fry.

Formal analysis: Baiming Zou, Fei Zou.

Investigation: Hudson P. Santos, Mike M. O'Shea, Rebecca C. Fry.

Methodology: Baiming Zou, Fei Zou.

Writing – original draft: Baiming Zou, Fei Zou.

Writing – review & editing: Baiming Zou, Hudson P. Santos, James G. Xenakis, Mike M. O'Shea, Rebecca C. Fry, Fei Zou.

References

1. Kinney MV et al. 15 million preterm births annually: what has changed this year? Reprod Health. 2012 9(28):1–4.
2. Bhutta A et al. Cognitive and behavioral outcomes of school-aged children who were born preterm: a meta-analysis. The Journal of the American Medical Association. 2002 288:728–737. https://doi.org/10.1001/jama.288.6.728
3. Anderson P and Doyle LW. Neurobehavioral outcomes of school-age children born extremely low birth weight or very preterm in the 1990s. The Journal of the American Medical Association. 2003 289:3264–3272. https://doi.org/10.1001/jama.289.24.3264
4. Moster D et al. Long-term medical and social consequences of preterm birth. N Engl J Med. 2008 359:262–273. https://doi.org/10.1056/NEJMoa0706475
5. Johnson S and Marlow N. Preterm Birth and Childhood Psychiatric Disorders. Pediatric Research. 2011 69(5):11R–18R. https://doi.org/10.1203/PDR.0b013e318212faa0
6. Raju T et al. Long-Term Healthcare Outcomes of Preterm Birth: An Executive Summary of a Conference Sponsored by the National Institutes of Health. The Journal of Pediatrics. 2016 181: 309–318. https://doi.org/10.1016/j.jpeds.2016.10.015
7. Johnson S and Marlow N. Early and long-term outcome of infants born extremely preterm. Arch Dis Child, 102, 97–102. https://doi.org/10.1136/archdischild-2015-309581
8. Risnes K., et al. (2021). Mortality Among Young Adults Born Preterm and Early Term in 4 Nordic Nations. JAMA Netw Open. 2017 4(1):e2032779. https://doi.org/10.1001/jamanetworkopen.2020.32779
9. Bangma JT et al. Assessing Positive Child Health among Individuals Born Extremely Preterm. J Pediatr. 2018 (202):44–49.
10. Bodnar LM and Simhan HN. The prevalence of preterm birth and season of conception. Paediatric and Perinatal Epidemiology. 2008 22:538–545. https://doi.org/10.1111/j.1365-3016.2008.00971.x
11. Pierrat V et al. Neurodevelopmental outcomes at age 5 among children born preterm: EPIPAGE-2 cohort study. BMJ. 2021 373:1–12.
12. Bangma JT et al. Early life antecedents of positive child health among 10-year-old children born extremely preterm. Pediatr Res. 2019 86(6):758–765. https://doi.org/10.1038/s41390-019-0404-x
13. Cho CR et al. The application of systems biology to drug discovery. Curr. Opin. Chem. Biol. 2006 10:294–302. https://doi.org/10.1016/j.cbpa.2006.06.025
14. Cisek K et al. The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. Nephrol. Dial. Transplant. 2016 31:2003–2011. https://doi.org/10.1093/ndt/gfv364
15. Duan N et al. A comparison of alternative models for the demand for medical care. Journal of Business and Economic Statistics. 1983 1:115–126. https://doi.org/10.1080/07350015.1983.10509330
16. Manning WG and Mullahy J. Estimating log models: to transform or not to transform? Journal of Health Economics. 2001 20(4):461–494. https://doi.org/10.1016/S0167-6296(01)00086-8
17. Buntin MB and Zaslavsky AM. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. J Health Econ. 2004 23(3):525–542. https://doi.org/10.1016/j.jhealeco.2003.10.005
18. Mihaylova B et al. Review of statistical methods for analysing healthcare resources and costs. Health Economics. 2011 20:897–916. https://doi.org/10.1002/hec.1653
19. Glick HA, Doshi JA, Sonnad SS, Polsky D. Economic Evaluation in Clinical Trials: Oxford University Press. 2014
20. Belotti F et al. twopm: Two-part models. The Stata Journal. 2015 15(1):3–20. https://doi.org/10.1177/1536867X1501500102
21. Olsen MK and Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. J Am Stat Association. 2001 96:730–745. https://doi.org/10.1198/016214501753168389
22. Farewell VT, et al. Two-Part Models and Related Regression Models for Longitudinal Data. Annu. Rev. Stat. Appl. 2017 4:283–315. https://doi.org/10.1146/annurev-statistics-060116-054131
23. Pinheiro JC & Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics. 1995 4:12–35. https://doi.org/10.2307/1390625
24. Tooze JA, et al. Analysis of repeated measures data with clumping at zero. Statistical Methods in Medical Research. 2002 11:341–355. https://doi.org/10.1191/0962280202sm291ra

25. Lai TL & Shih MC A Hybrid Estimator in Nonlinear and Generalised Linear Mixed Effects Models. Biometrika. 2003 90(4):859–879. https://doi.org/10.1093/biomet/90.4.859

26. O'Shea TM, Allred EN, Dammann O, et al. The ELGAN study of the brain and related disorders in extremely low gestational age newborns. Early Human Development. 2009 85(11):719–725. https://doi.org/10.1016/j.earlhumdev.2009.08.060

27. Taylor GL, O'Shea TM. Extreme prematurity: Risk and resiliency. Curr Probl Pediatr Adolesc Health Care. 2022 52(2):101132. https://doi.org/10.1016/j.cppeds.2022.101132

28. Kuban KC, Allred EN, O'Shea M, et al. An algorithm for identifying and classifying cerebral palsy in young children. J Pediatr. 2008 153(4):466–472. https://doi.org/10.1016/j.jpeds.2008.04.013

29. Wood CT, Linthavong O, Perrin EM, et al. Antecedents of Obesity Among Children Born Extremely Preterm. Pediatrics. 2018 142(5): e20180519. https://doi.org/10.1542/peds.2018-0519

30. Jackson WM, O'Shea TM, Allred EN, Laughon MM, Gower WA, Leviton A. Risk factors for chronic lung disease and asthma differ among children born extremely preterm. Pediatr Pulmonol. 2018 53 (11):1533–1540. https://doi.org/10.1002/ppul.24148

31. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika. 1988 75:383–386. https://doi.org/10.1093/biomet/75.2.383