A note on prognostic accuracy evaluation of regression models applied to longitudinal autocorrelated binary data

Barbati Giulia(1), Farcomeni Alessio(2), Pasqualetti Patrizio(3), Sinagra Gianfranco(1), Bovenzi Massimo(4)

ABSTRACT

BACKGROUND: Focus of this work was on evaluating the prognostic accuracy of two approaches for modelling binary longitudinal outcomes, a Generalized Estimating Equation (GEE) and a likelihood based method, Marginalized Transition Model (MTM), in which a transition model is combined with a marginal generalized linear model describing the average response as a function of measured predictors.

METHODS: A retrospective study on cardiovascular patients and a prospective study on sciatic pain were used to evaluate discrimination by computing the Area Under the Receiver-Operating-Characteristics curve, (AUC), the Integrated Discrimination Improvement (IDI) and the Net Reclassification Improvement (NRI) at different time occasions. Calibration was also evaluated. A simulation study was run in order to compare model's performance in a context of a perfect knowledge of the data generating mechanism.

RESULTS: Similar regression coefficients estimates and comparable calibration were obtained; an higher discrimination level for MTM was observed. No significant differences in calibration and MSE (Mean Square Error) emerged in the simulation study; MTM higher discrimination level was confirmed.

CONCLUSIONS: The choice of the regression approach should depend on the scientific question being addressed: whether the overall population-average and calibration are the objectives of interest, or the subject-specific patterns and discrimination. Moreover, some recently proposed discrimination indices are useful in evaluating predictive accuracy also in a context of longitudinal studies.

Key words: Area Under the Receiver-Operating-Characteristics curve (AUC); Net Reclassification Improvement (NRI); Integrated Discrimination Improvement (IDI); Generalized Estimating Equation (GEE); Marginalized Transition Model (MTM); longitudinal binary data.

(1) Cardiovascular Department, "Ospedali Riuniti" and University of Trieste, Italy

(2) Department of Public Health and Infectious Diseases, Sapienza - University of Rome, Rome, Italy

(3) Medical Statistics & Information Technology, Fatebenefratelli Association for Research, Isola Tiberina, Rome, Italy

(4) Clinical Unit of Occupational Medicine, Department of Medical Sciences, University of Trieste, Italy

CORRESPONDING AUTHOR: Giulia Barbati, PhD - Cardiovascular Dept; Polo Cardiologico, Ospedale di Cattinara; University of Trieste - Via Valdoni 1, 34149 Trieste, Italy - tel: +39 040 399 4198 - E-mail: giulia_barbati@yahoo.it

*DOI: **10.2427/10003***

Accepted on 6 August 2014; Published as Online First on 17 December, 2014

INTRODUCTION

Both in retrospective and prospective observational studies, repeated measurements are often taken over time to evaluate longitudinal dynamics of an outcome and factors affecting this evolution. In some instances, outcomes are binary indicators of presence of a particular condition. Statistical methods for analysis of longitudinal binary data have been rapidly developed during recent years; Diggle et al. provide the detailed review [1]. Either subject-specific and population-average estimates could be a major goal in this context. The data that motivated this manuscript come from two different research fields: a cardiovascular (CV) cohort of 503 idiopathic dilated cardiomyopathy (DCM) patients, enrolled from 1988 to 2007 in the Heart Muscle Disease Registry of Trieste, a database of a tertiary referral center on cardiomyopathies [2] and a database from an occupational medicine department (OM). The latter consisted in a prospective cohort study on 537 male professional drivers conducted over a two-year period (2004-2006) to investigate the relation of sciatic pain to measures of internal spinal load [3]. In CV patients, timing of implanted cardioverter defibrillator (ICD) for primary prevention of sudden death (SD) is, to this day, not clearly established: waiting some time could reduce useless implants but also increase the exposure to SD, and no guidelines are available. General criteria for ICD eligibility in DCM patients are Left Ventricular Ejection Fraction (LVEF) ≤0.35 and NYHA (New York Heart Association) functional classes II or III ("SCD-HeFT criteria": Sudden Cardiac Death Heart Failure Trial, [4]). The aim of this retrospective cohort study was to explore the longitudinal evolution of the presence/absence of SCD-HeFT criteria after therapy optimization in order to identify the optimal waiting time before implantation. In the OM prospective study, measures of daily whole body vibration (WBV) exposure and spinal load were derived from the biodynamic and epidemiological databases implemented by the German and Italian arms of the EU VIBRISKS project [5-6]. The main goal was on determining the boundary values for the internal lumbar load to prevent the occurrence of sciatic pain from a public-health intervention point of view.

The performance of the two following regression approaches were compared on the above datasets: a Generalized Estimating Equation, GEE method [7], and a likelihood based method, Marginalized Transition Models, MTM [8]. The former, adopts a semiparametric model in which only the marginal mean and the correlation of repeated measurements are specified. In the latter, a transition model that characterizes serial dependence is combined with a marginal generalized linear model that describes the average response as a function of measured predictors. Models' performances were assessed by calibration and discrimination. Calibration addresses the question of how closely the model-based risk estimates align with the observed outcomes. Discrimination focuses on a model's ability to distinguish between subjects who will (or did) develop the event of interest from those who will (did) not. A simulation study was also run, in order to compare performance of the above regression models in a context of a perfect knowledge of the true data generating mechanism.

METHODS

Real data modelling

We observe serial binary response data $Y_i = (Y_{i1},\ldots,Y_{in_i})$ on subjects $i=1,\ldots,n$ at times $t=1,\ldots,n_i$. In CV data, Y_{it} is a binary indicator of the SCD-HeFT condition at time t for subject i; in OM data, Y_{it} is a binary indicator of sciatic pain at time t for subiect i. We also observe a set of constant or time-varying covariates, respectively $Z_i = (Z_{i1},\ldots,Z_{ik})$ and $X_{it} = (X_{it,1},\ldots,X_{it,r})$ recorded for each subject at baseline and at each occasion. In CV data, a constant covariate indicates the initial condition of the subject, *'InitialCond'*, coded as an ordinal variable: 1=No SCD-HeFT at baseline, 2= SCD-HeFT at baseline, 3=NYHA IV at baseline, based on the increasing severity of the starting condition, and a time effect (*'Time'*) that assumes values from 0 to 48 (time scale is in months), depending on the number of observations per subject. In OM data, a constant covariate indicates age at entry in the study (*'Age'*); time-varying continuous covariates are: years of exposure at *"whole-body vibration"*

(*'WBV'*) in decades and measures of internal spinal load as the daily compressive dose *'Sed'* (in MPa), and a risk factor *'R'* (non-dimensional units) that measures the adverse health effects related to the compressive dose, both measured at six lumbar spine levels [3,9]. There are also time-varying categorical covariates that represents individual and work-related risk factors: physical work load (*'Posture'*, with four levels: mild, moderate, hard, very hard) and psychosocial work environment (*'Psycho'*, with four levels: good, acceptable, a little bit bad, bad); finally, the time effect (*'Time'*, with three occasions corresponding to surveys).

To obtain estimates for the regression of Y_{it} on covariates $E(Y_{it} \mid \mathbf{Z}_i;\mathbf{X}_{it})$ we assume that the regression model properly specifies the full covariate conditional mean defined as

$$\mu_{it} = E(Y_{it} \mid \mathbf{Z}_i;\mathbf{X}_{it}) = E(Y_{it} \mid Z_{i1},\dots,Z_{ik};X_{it,1},\dots,X_{it,r})$$

Since the outcome is a binary condition, the marginal generalized linear model specifies: $g(\mu_{it})=X_{it}\beta+Z_i\lambda$ where $g()$ is the logit link function $g(\mu_{it}) = \log \frac{\mu_{it}}{1\ \mu_{it}}$ and the regression coefficients β,λ measure the influence of covariates on the average response. We have to take into account the within-subject correlation that naturally arise from repeated observations on the same patients. For CV data, the following two regression models were estimated:

(CV.a) :

$$g(\mu_{it}) = \beta_0 + \beta_1 * Time$$

(CV.b):

$$g(\mu_{it}) = \beta_0 + \beta_1 * Time + \lambda_1 * InitialCond$$

Two regression models were estimated for OM data, in order to compare the two different measures of spinal load available:

(OM.a):

$$g(\mu_{it}) = \beta_0 + \beta_1 * Time + \lambda_1 * Age + \beta_2 * WBW + \beta_3 * Posture + \beta_4 * Psycho + \beta_5 * Sed_{L5S1}$$

(OM.b):

$$g(\mu_{it}) = \beta_0 + \beta_1 * Time + \beta_2 * Posture + \beta_3 * Psycho + \beta_4 * R_{L5S1}$$

In model (OM.b) covariates Age and WBW were excluded since they are both present in the computation of the risk factor R_{L5S1}. The choice of the spinal levels L5-S1 for the measures of internal spinal load was based on a preliminary analysis among all six spinal levels that indicated Sed_{L5S1} and R_{L5S1} as the most representative of the extremely correlated group of spinal levels linearly affecting the outcome. Each of the above models was estimated with the two different regression approaches: the GEE method [7,10] where specification of a working correlation structure is required to take account of the correlation and main choices are among independence, exchangeable, autoregressive or stationary. In the present study, the autoregressive of order one (AR1) working correlation structure was adopted, in analogy with the second regression approach, MTM. These are a general parametric class of serial dependence models that permit likelihood based marginal regression analysis of binary response data [1,8] where covariate effects and within-subject association are modelled through a single equation: given the immediate previous response in a first-order Markov model [11] the current response variable is assumed to be conditionally independent of any previous outcome variables $E(Y_{it} \mid Y_{ij}, j<t) = E(Y_{it} \mid Y_{it-1})$. The probabilities that define the first order Markov process are given by $p_{it,0} = E(Y_{it} \mid Y_{it-1}=0)$ and $p_{it,1} = E(Y_{it} \mid Y_{it-1}=1)$, i.e. the probability to observe an outcome at time *t* if the outcome condition was absent or present in the previous follow up. The first-order MTM adopted in the present study is specified by assuming a regression structure for the marginal mean, using the above defined generalized linear model $g(\mu_{it})$, and this model is constrained by the transition probabilities to satisfy $\mu_{it} = p_{it,1}\mu_{it-1} + p_{it,0}(1\ \ \mu_{it-1})$. A parameter of serial dependence of Y_{it} on Y_{it-1} is estimated that measures the log odds to have the outcome at time t among subjects who had the outcome at time *t-1* compared to subjects who had not: $\alpha = log(OR(Y_{it}, Y_{i,t-1}))$ where *OR* stands for odds ratio. R packages 'geepack' [12] and 'mtm' were used to fit the models.

Prognostic accuracy evaluation

The relative behaviour of GEE with respect to likelihood based models had already been exploited in various context of study design and inference objectives, with the help of simulation studies and real data applications [13-14]; but specific considerations about discrimination and calibration have not been particularly explored in the literature. To evaluate discrimination, i.e. the model's ability to distinguish between subjects who will develop the event or condition of interest from those who will not, three discrimination indices were computed. First of all, the Area Under the Receiver-Operating-Characteristics curve, AUC: as it is well known [15], AUC represents the probability that given two subjects, one who will develop an event and the other who will not, the model will assign a higher probability of an event to the former. To take into account the longitudinal structure of the data, AUCs were computed separately at the different times for the two approaches, where *Model*$_x$=GEE or MTM, *t*=time:

$$AUC_t^{Model_x} = P(\hat{p}_t^i(Model_x) > \hat{p}_t^j(Model_x) | Y_{it} = 1, Y_{jt} = 0)$$

where $\hat{p}_t^i$ is the probability of event estimated by the model for subject *i* at time *t*. AUCs values range from 0.50 (useless) to 1 (perfect discrimination). The De Long test was calculated to evaluate differences among ROC curves [16].

The recently introduced [17] Net Reclassification Index (NRI) and the Integrated Discrimination Improvement (IDI) were also computed following the longitudinal structure of the data. NRI evaluates the relative increase in the predicted probabilities for subjects who experience the event and the decrease for subjects who do not. NRI is the sum of two components, 'NRI for events' and 'NRI for non events': if we indicate the MTM model-based vector of predicted probabilities of event at time *t* by $\hat{p}^t_{MTM}$ and the GEE model-based vector of probabilities by $\hat{p}^t_{GEE}$, NRI at time t was estimated as:

$$NRI_t = \frac{\sum_{i\ in\ events_t}(\hat{p}^t_{MTM}(i) - \hat{p}^t_{GEE}(i))}{\#events_t} - \frac{\sum_{j\ in\ non\ events_t}(\hat{p}^t_{MTM}(j) - \hat{p}^t_{GEE}(j))}{\#non\ events_t}$$

where the first term ('NRI for events') quantifies improvement in sensitivity and the negative of the second term ('NRI for non events') quantifies improvement in specificity. Heuristic benchmarks to interpreting NRI values are: NRI around 0.20 indicates a weak improvement; around 0.40 a medium improvement, and greater than 0.60 a strong improvement [18].

IDI is the difference in discrimination slopes, i.e. the difference between mean predicted probabilities of an event for those with events and the corresponding mean for those without events; thus, it jointly quantifies the overall improvement in sensitivity and specificity, and it can be estimated at time *t* as:

$$IDI_t = (\bar{\hat{p}}^t_{MTM,events} - \bar{\hat{p}}^t_{MTM,non\ events}) - (\bar{\hat{p}}^t_{GEE,events} - \bar{\hat{p}}^t_{GEE,non\ events})$$

where $\bar{\hat{p}}^t_{Model_x,events}$ is the mean predicted probability of *Model*$_x$ (GEE or MTM) at time *t* in the group of subject with events, and $\bar{\hat{p}}^t_{Model_x,non\ events}$ is the mean predicted probability in the group of subject without events. Values of IDI around 0.006 indicates a weak improvement, around 0.04 a medium improvement and around 0.10 a strong improvement [18]. For NRI and IDI asymptotic tests for the null hypothesis of no difference between models can be applied, and 95% confidence intervals have been computed. 'Calibration at large' was also evaluated, calculating means of the predicted risks by the two approaches at different times while comparing them with the observed outcomes.

Simulation study

In order to complement the illustration from the case studies we also ran a brief simulation. We fixed regression parameters β and generated three covariates. The first, associated with a negative slope, recorded time from baseline measurement, the second was a time-varying indicator generated from a fair Bernoulli trial, and the last one a time-fixed continuous covariate generated from a standard Gaussian. In order to sample the outcome we adopted the approach described in [19], which is based on the marginal means arising from the above described fixed effects and covariates. A first-order autocorrelation (AR1) structure was used. Hence, we did not generate data from any of the two models afterwards used for fitting. We fixed T=4 time occasions and repeated data generation, model estimation and evaluation of the results for different values of the sample size n (250,500,1000), and correlation coefficient rho (0.1,0.25,0.33,0.5). The results are averaged over 1000 replicates.

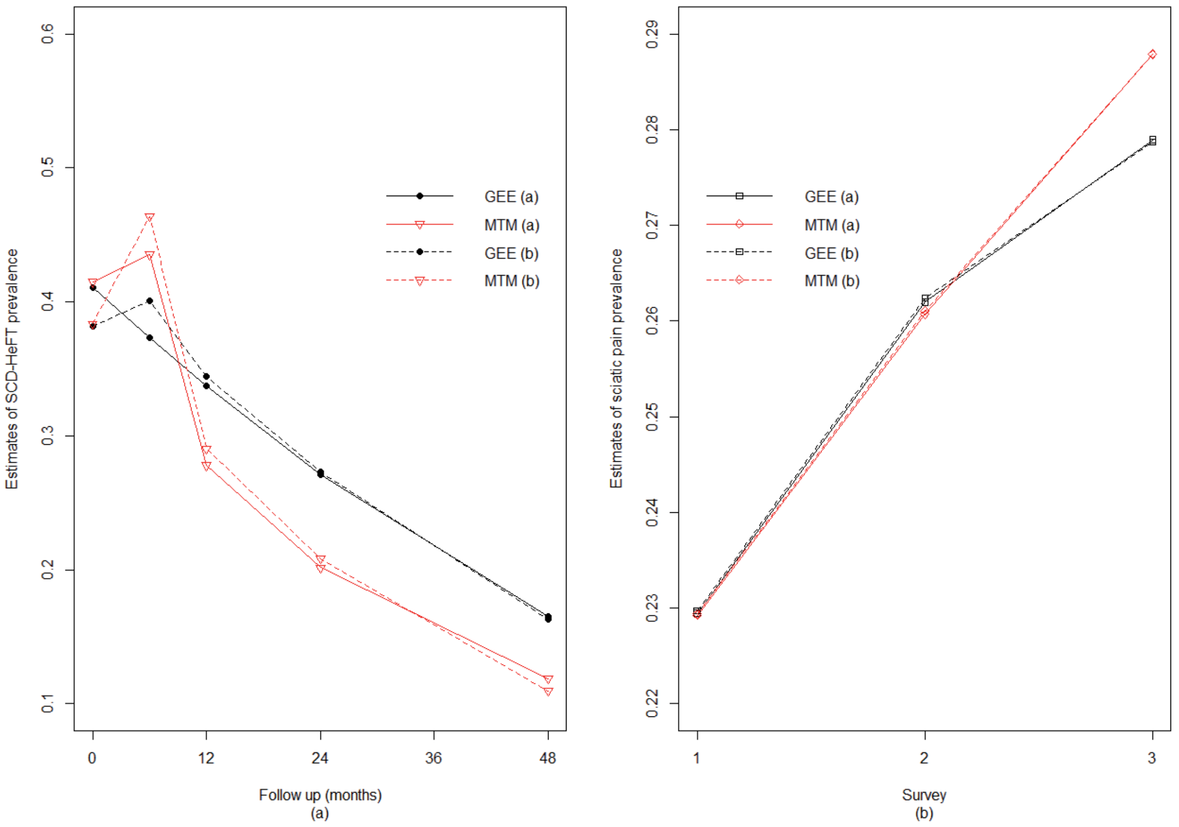
Results from case studies

CV data

At diagnosis and at each follow-up study, population was divided into two groups by the binary condition "SCD-HeFT criteria" yes or no. The first four evaluations were considered, on average at 6 (range 3-9), 12 (9-18), 24 (18-36) and 48 (36-60) months after diagnosis. Patients with all measures were defined as *"Complete Cases"* (CC). Patients with some follow-up visits but not others were defined as *"Intermittent Drop-out"* (ID); *'Drop-out at Death'* (DD) were those fully observed until death, and finally *"Baseline Drop-out"* (BD) were those that drop-out after baseline, but were known to be alive after 48 months. CC group counted 165 (33%) patients; ID were 221 (44%); 47 patients (9%) were DD, and 70 (14%) were BD. Independently from the number of patients included (from all patients until 'CC') the rate of temporal decline in the log-odds of SCD-HeFT estimated by GEE and MTM models by using (CV.a) was comprised between -0.02 to -0.03 per time increment (odds ratio between 0.97 and 0.98 per month, nearly 0.70-0.76 in 12 months). Following (CV.b),

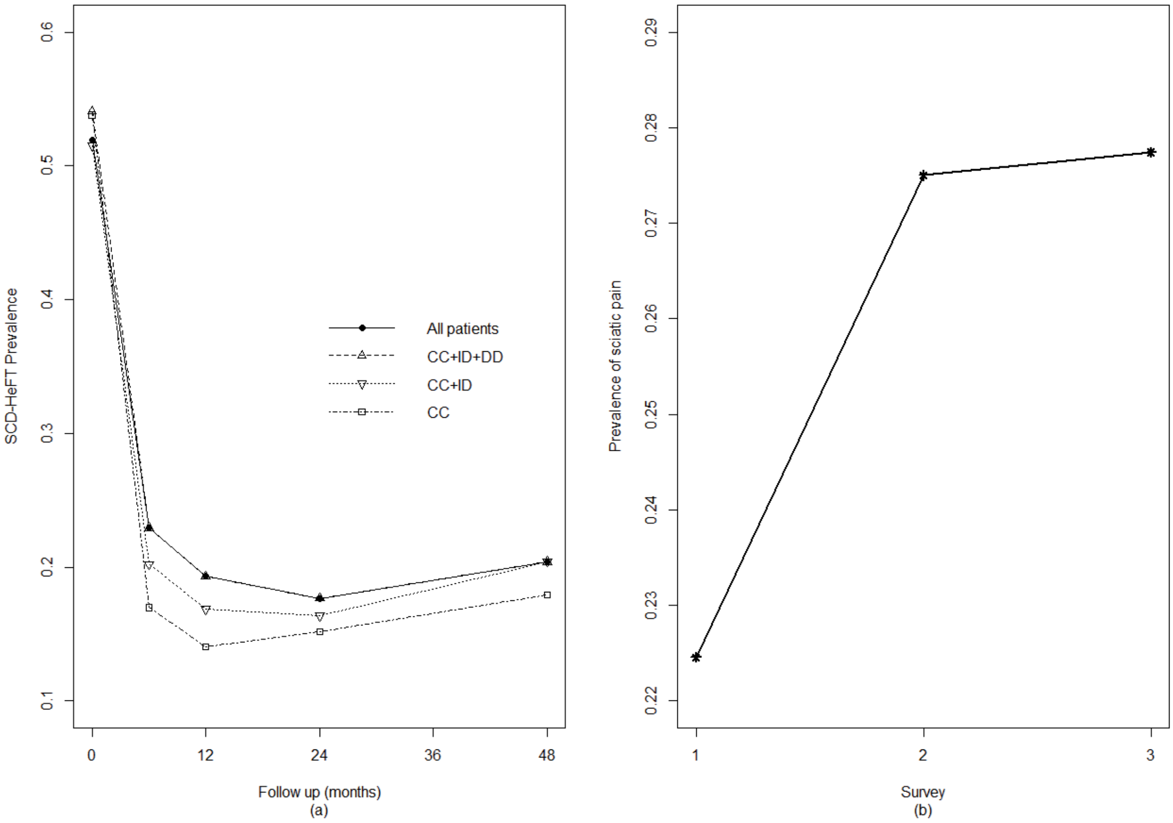
FIGURE 1

(a): ESTIMATED PREVALENCE OF SCD-HEFT CRITERIA, CV DATA.
(b): ESTIMATED PREVALENCE OF SCIATIC PAIN, OM DATA.



SUPPLEMENTARY FIGURE 1

(a): OBSERVED PREVALENCE OF SCD-HEFT CRITERIA, CV DATA.
(b) OBSERVED PREVALENCE OF SCIATIC PAIN, OM DATA.



the initial condition was highly relevant in increasing the probability to be or not SCD-HeFT at successive follow-up (log-odds from 1.5 to 2.4, depending on the subgroups used, table 1). The MTM serial dependence of first-order was always significant and strong: the log odds to have SCD-HeFT at time *t* among subjects who had SCD-HeFT at time *t-1* with respect to subjects who had not was estimated between 1.7 and 2.2. The AR1 within-subject correlation estimated by GEE was also significant and strong for both models (between 0.49 and 0.51). Considering all patients, the observed prevalence of SCD-HeFT at baseline was 52% , that sharply decreased at 23% at 6 months, and remained quite stable until 48 months (respectively at 12 and 24 months: 19% and 18%, and 20% at 48 months); the others sub-groups of patients showed very similar trends (Supplementary Figure 1(a)). Calibration at large, expressed as estimates of prevalence at different times, was biased towards lower values at baseline and higher values at 6-12 months, both for models (CV.a) and (CV.b) and for both regression approaches; more precise estimates at the end of follow up were obtained, at 24 and 48 months. In particular, at baseline the estimated prevalence was 41% for GEE and 42% for MTM (both in models CV.a and CV.b); it decreased respectively at 34% and 28% at 12 months and at 17% and 13% at 48 months, showing an higher precision of GEE at the end of follow up and lower in the middle (Figure 1 (a)).

OM data

Characteristics of the study population have been already extensively described in various papers [4,21]. Briefly, the cohort included male professional drivers (n=628) employed in several industries and public utilities located in various Provinces of Italy. The rate of participation in the initial cross-sectional survey was 95.2% (n=598); five hundred and thirty-seven responders participated in the successive follow up surveys over the calendar periods 2004-2006. Causes for loss to the follow-up

TABLE 1

CV DATA: GEE AND MTM RESULTS: ESTIMATE (SE)

		ALL PATIENTS	CC+ID+DD	CC+ID	CC
GEE (a)	Intercept	-0.360 (0.078)	-0.348 (0.083)	-0.485 (0.088)	-0.614 (0.121)
	Time	-0.026 (0.004)	-0.027 (0.004)	-0.024 (0.004)	-0.023 (0.005)
	Alpha (AR1 corr)	0.497 (0.043)	0.499 (0.044)	0.506 (0.050)	0.514 (0.070)
		ALL PATIENTS	CC+ID+DD	CC+ID	CC
GEE (b)	Intercept	-4.360 (0.388)	-3.933 (0.372)	-3.975 (0.390)	-3.227 (0.523)
	Time	-0.032 (0.005)	-0.031 (0.005)	-0.029 (0.005)	-0.026 (0.006)
	Initial Condition	2.421 (0.240)	2.150 (0.223)	2.108 (0.241)	1.561 (0.321)
	Alpha (AR1 corr)	0.503 (0.201)	0.466 (0.160)	0.477 (0.172)	0.494 (0.131)
		ALL PATIENTS	CC+ID+DD	CC+ID	CC
MTM (a)	Intercept	-0.345 (0.084)	-0.330 (0.089)	-0.469 (0.095)	-0.600 (0.142)
	Time	-0.029 (0.004)	-0.029 (0.004)	-0.027 (0.004)	-0.026 (0.006)
	Alpha (serial dep.)	2.006 (0.159)	2.009 (0.159)	2.051 (0.165)	2.208 (0.221)
		ALL PATIENTS	CC+ID+DD	CC+ID	CC
MTM (b)	Intercept	-3.896 (0.278)	-3.634 (0.283)	-3.711 (0.296)	-3.181 (0.418)
	Time	-0.036 (0.005)	-0.034 (0.005)	-0.032 (0.005)	-0.029 (0.007)
	Initial Condition	2.153 (0.154)	1.977 (0.155)	1.954 (0.163)	1.543 (0.228)
	Alpha (serial dep.)	1.742 (0.181)	1.705 (0.179)	1.744 (0.185)	2.052 (0.235)

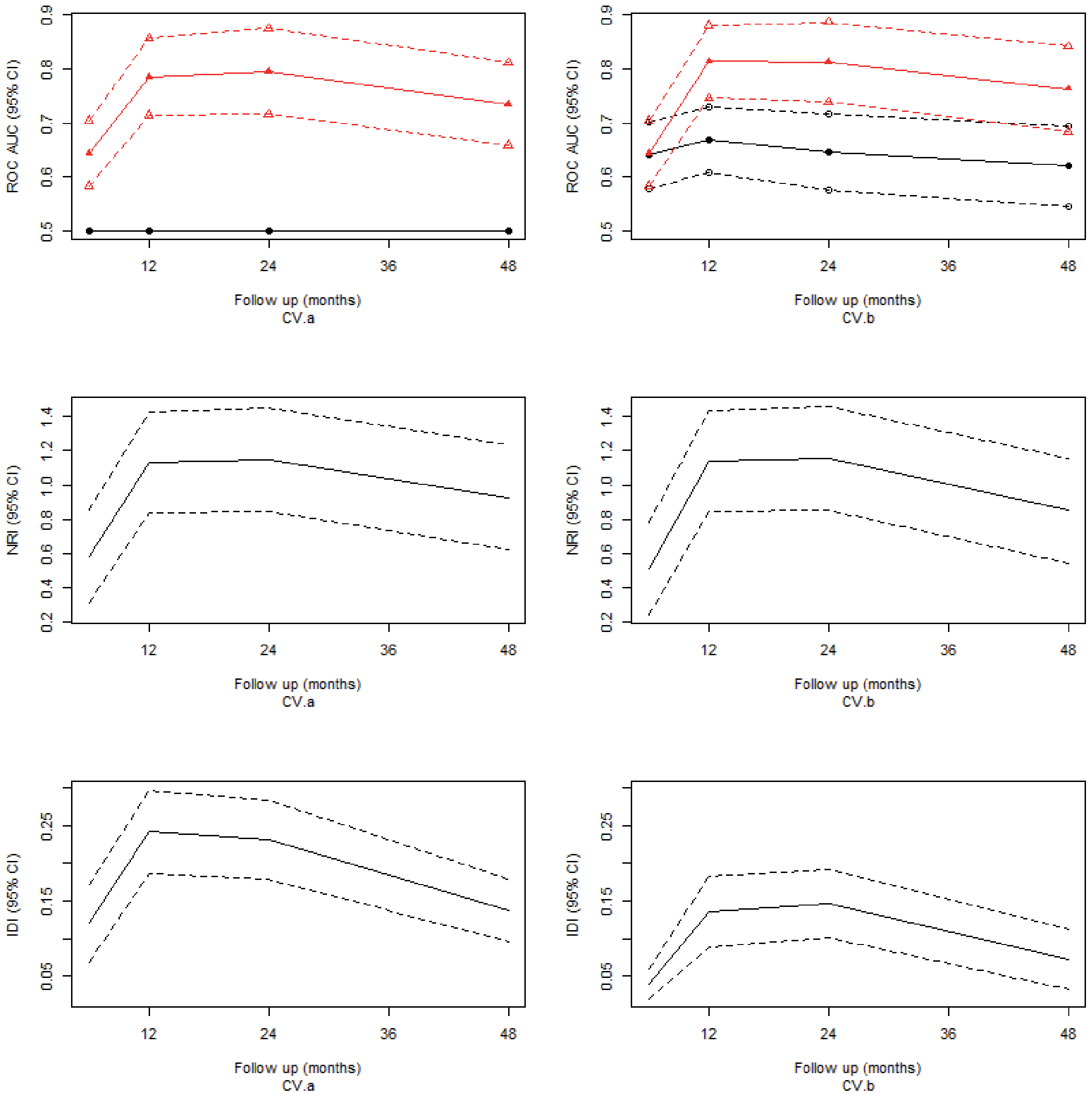
were change of residence (n=15), consent refusal (n=36), and undetermined reasons (n=10). Not all 537 subjects participated in all the three surveys: 80% in the first, 89% in the second and 83% in the third; all subjects had at least two measures, but only 59% of them had also a third measure, for a total of 1391 observations. Moreover, some subjects had missing values for several covariates included in the regression models, therefore the total number of observations included in the multivariable models was 1359. At the first survey, the mean age at entry was 41 (sd=8), the average years of exposure at *"whole-body vibration"* (*'WBV'*) in decades was 1.7 (sd=1), the daily compressive dose Sed_{L5S1} averaged at 0.29 (sd=0.1) MPa, and the risk factor at 0.27 (0.13) units. Physical work load (*'Posture'*) was *"hard"* for the 20% of subjects and "very hard" for the 17%; psychosocial work environment (*'Psycho'*) was *"bad"* for 15% of subjects. All these time-varying factors remained quite stable during successive surveys, except for an increase observed in the percentage of the *"very hard"* physical work load (increased at 27-28% in the second and third survey).

For model (OM.a), regression coefficients of GEE and MTM were very similar: an OR of 1.3 for a unit increase of WBW was estimated, postures *'hard'* and *'very hard'* were significantly affecting the outcome with respect to the level reference *'mild'*, as well as psychosocial work environment defined as *'bad'* showed an OR between 1.5 and 1.6 with respect to the level reference 'good'. For a change of 0.1 MPa in Sed_{L5S1} the estimated OR was 1.2. Also for model (OM.b), regression estimates for GEE and MTM were nearly equivalents: time become borderline significant, with an OR of 1.12, postures *'hard'* and *'very hard'* were still significantly affecting the outcome with respect to the level reference, instead the psychosocial work environment was no longer significant. For a change of 0.1 units in R_{L5SL} the estimated OR was 1.3. The AR1 within-subject correlation estimated by GEE was always significant, at a level of 0.5. The MTM serial dependence of first-order pointed strongly to serial dependence: the log odds to have sciatic pain at time *t* among subjects who had sciatic pain at time *t-1* with respect to subjects who had not was estimated at 2.2 (table 2).

At the first survey, the observed percentage of subjects with sciatic pain was 22.5% and it increased at 27.5% in the second survey and at 27.7% in the third (Supplementary Figure 1 (b)). Both for models (OM.a) and (OM.b) calibration of GEE and MTM was satisfactory: mean predicted probabilities were

FIGURE 2

TOP-DOWN: AUCS (SOLID LINES) AND CORRESPONDING 95% CI (DASHED LINES);
IN RED AND TRIANGLES MTM VALUES, IN BLACK AND DOTS GEE VALUES, FOR CV.A AND CV.B.
NRI VALUES (SOLID LINES) AND CORRESPONDING 95% CI (DASHED LINES) FOR CV.A AND CV.B.
IDI VALUES (SOLID LINES) AND CORRESPONDING 95% CI (DASHED LINES) FOR CV.A AND CV.B.



respectively 23% and 23% at the 1st, 26.2% and 26.1% at the 2nd, 27.9% and 28.7% at the 3rd survey (Figure 1 (b)). Both methods showed a similar trend for the estimated prevalence, over-estimating at the 1st and 3rd occasion and slightly under-estimating at the 2nd survey. An higher precision of GEE was observed in the 3rd survey, at the end of the study.

Discrimination

CV data

Considering all patients, the discrimination level of MTM was always significantly superior to GEE: AUCs for MTM computed at different follow-up times ranged from 0.65 to 0.80 following (CV.a) and nearly the same following (CV.b). Of note, for (CV.a), AUCs computation are useless for GEE, since GEE predicted probabilities have the same value for all patients, being *'Time'* the only covariate in the regression model (and it is at a fixed value computing AUCs separately at different follow-up times). For (CV.b), AUCs values for GEE were always around 0.65, significantly lower with respect to MTM (De Long test p values always <=0.001, except at 6 months). NRI values showed more or less the same trend of AUCs: starting with 'lower' (but high in the NRI interpretation scale) values at 6 months (around 0.60 for both CV.a and CV.b), then

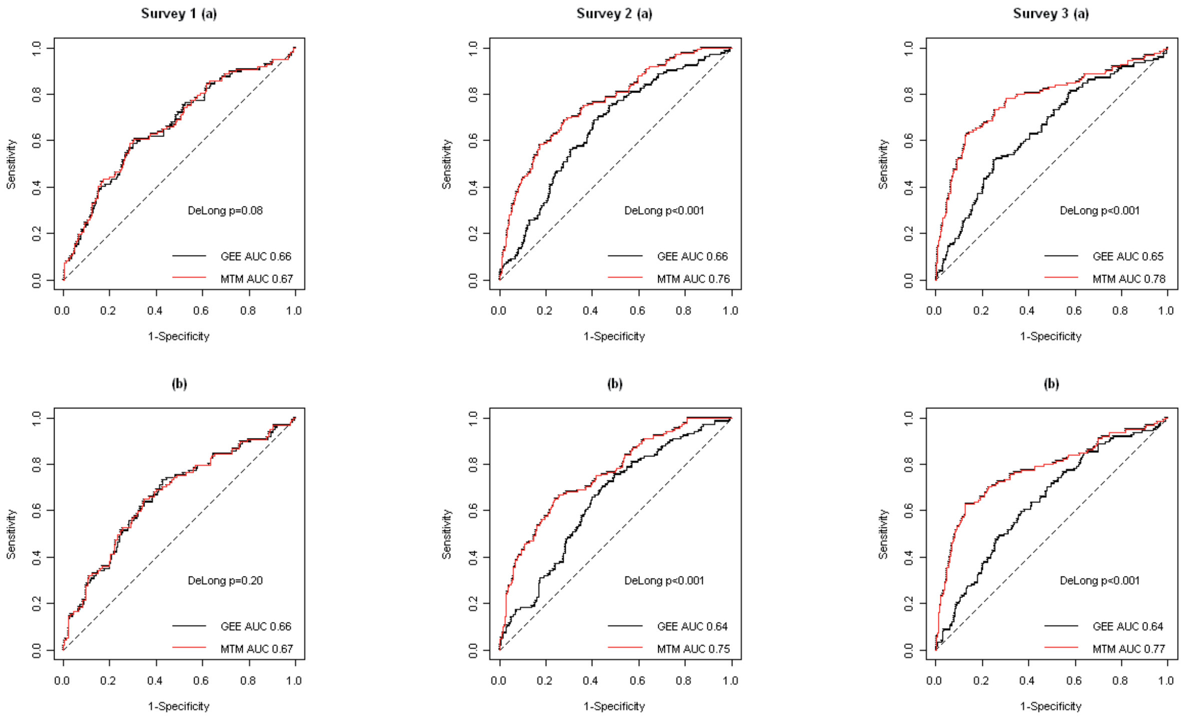
TABLE 2

OM DATA: GEE AND MTM RESULTS: ESTIMATE (SE). IN ITALIC NOT SIGNIFICANT ESTIMATES

	GEE (A)		GEE (B)
Intercept	-2.461 (0.592)	Intercept	-2.339 (0.483)
Time	0.102 (0.063)	Time	0.118 (0.062)
Age	-0.004 (0.014)	-	-
WBW	0.267 (0.120)	-	-
Posture: moderate	0.216 (0.170)	Posture: moderate	0.222 (0.170)
hard	0.443 (0.164)	hard	0.457 (0.163)
very hard	0.599 (0.169)	very hard	0.615 (0.168)
Psycho: acceptable	0.009 (0.201)	Psycho: acceptable	-0.049 (0.199)
little bad	-0.049 (0.194)	little bad	-0.167 (0.197)
bad	0.438 (0.223)	bad	0.261 (0.230)
SED L5 S1	0.181 (0.079)	R L5 S1	0.260 (0.081)
Alpha (AR1 corr)	0.524 (0.060)	Alpha (AR1 corr)	0.513 (0.056)
	MTM (A)		MTM (B)
Intercept	-2.489 (0.574)	Intercept	-2.359 (0.482)
Time	0.101 (0.072)	Time	0.116 (0.071)
Age	-0.004 (0.014)	-	
WBW	0.262 (0.121)	-	
Posture: moderate	0.265 (0.170)	Posture: moderate	0.265 (0.169)
Hard	0.467 (0.212)	Hard	0.483 (0.179)
Very hard	0.663 (0.178)	Very hard	0.674 (0.178)
Psycho: acceptable	-0.015 (0.192)	Psycho: acceptable	-0.077 (0.191)
Little bad	-0.074 (0.184)	Little bad	-0.202 (0.188)
Bad	0.471 (0.179)	Bad	0.265 (0.219)
SED L5 S1	0.177 (0.079)	R L5 S1	0.260 (0.078)
Alpha (serial dep.)	2.186 (0.191)	Alpha (serial dep.)	2.184 (0.191)

FIGURE 3

ROC AUCS ESTIMATED AT DIFFERENT SURVEY FOR MODEL OM.a (UP) AND OM.B (DOWN) FOR THE TWO REGRESSION APPROACHES.



increasing and peaking between 12 and 24 months (from 1.18 to 1.23 respectively for CV.a and CV.b) and then decreasing at 48 months (from 0.91 to 0.62, CV.a and CV.b). NRI indices were always significant, except at 48 months for model (CV.b) (figure 2 and supplementary table 1). IDI showed a different behaviour in the two regression models: in case of (CV.a), IDI values were higher than in case of (CV.b), indicating that an improvement both in sensibility and in specificity was achieved by GEE using (CV.b), even if showing significantly lower values with respect to MTM (figure 2 and supplementary table 1). Results for the other groups of patients were in line with ‘all population’ (supplementary table 1).

OM data

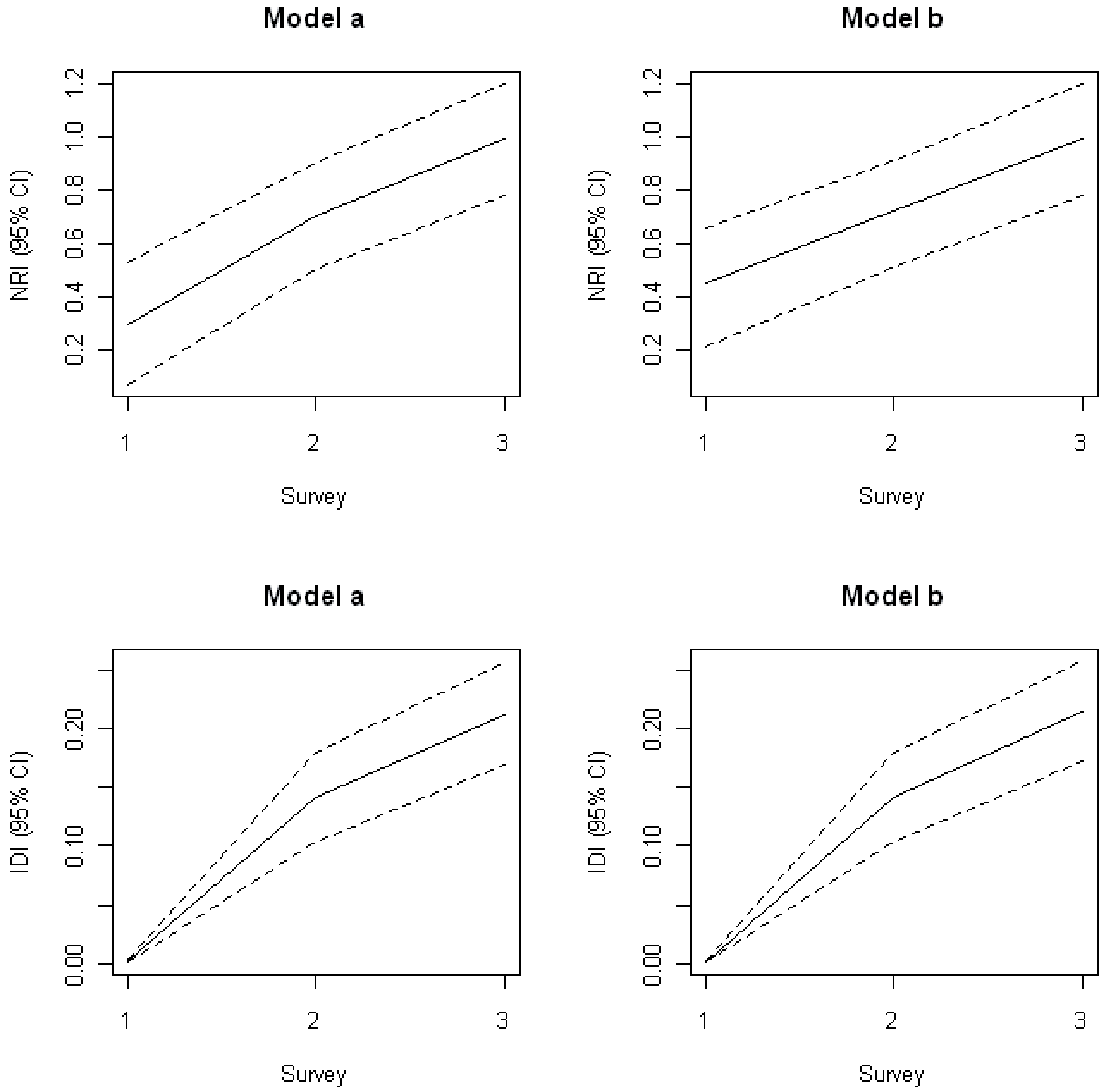
Except for the first survey, where no significant differences were observed, the discrimination level of MTM was always significantly superior to GEE: AUCs for MTM computed at the 2nd and 3rd survey ranged from 0.76 to 0.78, following (OM.a) and (OM.b), with respect to the GEE AUCs range of 0.64-0.67 (figure 3 and supplementary table 2, De Long p<0.001). NRI values were always significant and showed the same trend both for model (OM.a) and (OM.b): a nearly linear increase between the first and the last survey (figure 4 and supplementary table 2). The time trend for IDI values was even more impressive, starting with low values at the first survey (0.002-0.003) and increasing at 0.140-0.200 at successive surveys (figure 4 and supplementary table 2). No relevant differences between models (OM.a) and (OM.b) were observed, indicating a similar role of the two different measures of internal spinal load available.

Simulation results

With respect to the estimated MSE (Mean Square Error) and in line with expectations, a decreasing trend was observed for increasing sample size. The two methods yielded approximately

FIGURE 4

NRI AND IDI VALUES (SOLID LINES) AND CORRESPONDING 95% CI (DASHED LINES) ESTIMATED FOR OM.a (LEFT) AND OM.b (RIGHT)



the same MSE, which is also seen to be constant with respect to the level of autocorrelation (Figure 5 panel (a)). On the other hand, AUCs, NRI and IDI were quite insensitive to sample size (Table 3). In Figure 5 we report results corresponding to n=500 for visual illustration. Both Table 3 and Figure 5 show that when the autocorrelation is low (rho=0.1), MTM is nearly equivalent to GEE. As the autocorrelation grows, the predictive accuracy of MTM gets better and better than GEE in terms of AUCs, NRI and IDI. A stable pattern is seen across time occasions, where on the first occasion the two methods perform similarly, and then they tend to lead to different predictions at later times. Calibration results were nearly the same for the two methods.

DISCUSSION

Analysis of longitudinal data is challenging due to the need to address correlated outcomes, missing data and time-varying exposure processes. In this paper we focused on the ways to evaluate prognostic

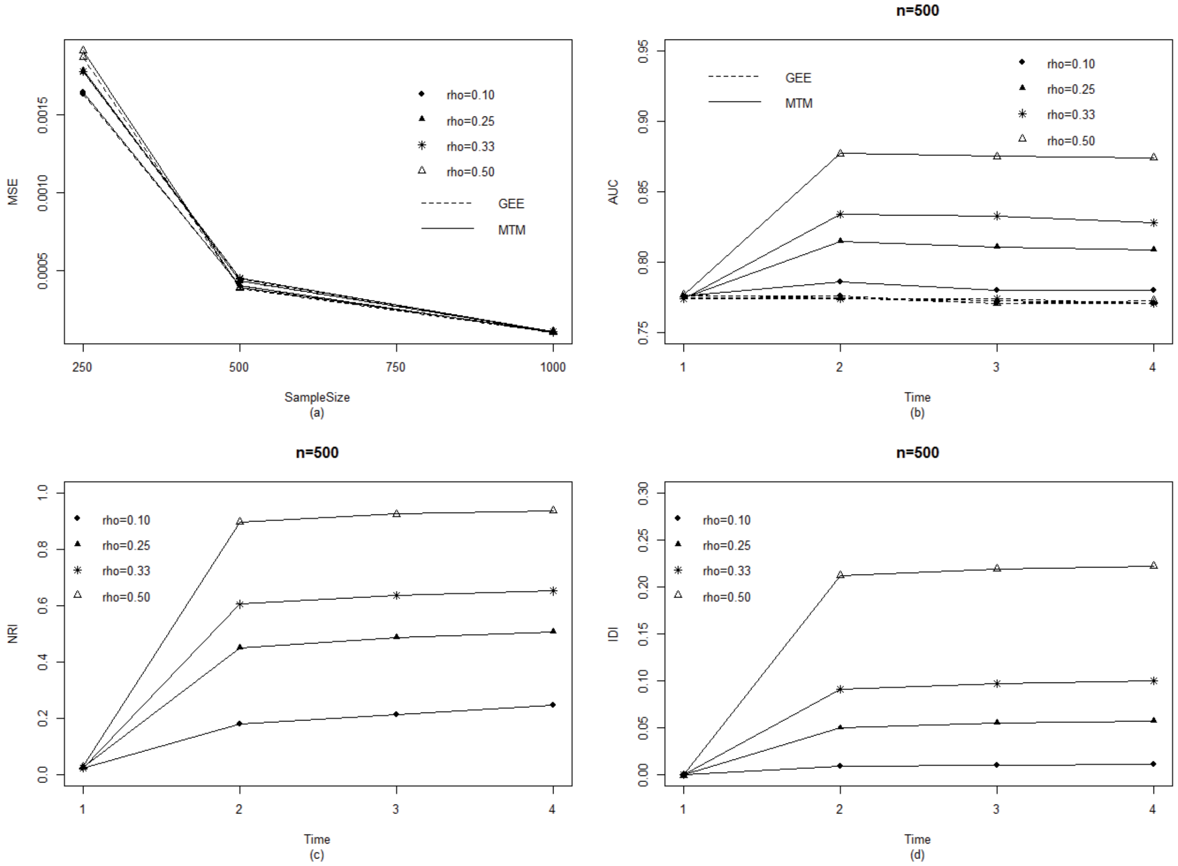
TABLE 3

SIMULATION STUDY RESULTS

SAMPLE SIZE	RHO	TIME	AUC.GEE	AUC.MTM	NRI	IDI
250	0.1	1	0,78	0,78	0,05	0,0000
		2	0,78	0,79	0,15	0,0100
		3	0,77	0,78	0,20	0,0110
		4	0,78	0,78	0,22	0,0120
	0.25	1	0,78	0,78	0,04	0,0000
		2	0,78	0,82	0,44	0,0540
		3	0,77	0,81	0,47	0,0580
		4	0,77	0,81	0,50	0,0610
	0.33	1	0,78	0,78	0,09	0,0000
		2	0,78	0,84	0,58	0,0920
		3	0,77	0,83	0,62	0,0990
		4	0,77	0,83	0,64	0,1020
	0.50	1	0,77	0,77	0,02	0,0000
		2	0,77	0,88	0,88	0,2150
		3	0,77	0,87	0,91	0,2200
		4	0,77	0,87	0,91	0,2210
500	0.1	1	0,78	0,78	0,02	0,0000
		2	0,78	0,79	0,18	0,0090
		3	0,77	0,78	0,21	0,0100
		4	0,77	0,78	0,25	0,0110
	0.25	1	0,78	0,78	0,03	0,0000
		2	0,78	0,82	0,45	0,0500
		3	0,77	0,81	0,49	0,0550
		4	0,77	0,81	0,51	0,0570
	0.33	1	0,77	0,77	0,02	0,0000
		2	0,77	0,83	0,61	0,0910
		3	0,77	0,83	0,64	0,0970
		4	0,77	0,83	0,65	0,1000
	0.50	1	0,78	0,78	0,03	0,0000
		2	0,78	0,88	0,90	0,2120
		3	0,77	0,88	0,93	0,2190
		4	0,77	0,87	0,94	0,2220
1000	0.1	1	0,77	0,77	0,01	0,0000
		2	0,78	0,79	0,21	0,0090
		3	0,77	0,78	0,25	0,0090
		4	0,77	0,78	0,27	0,0100
	0.25	1	0,78	0,78	0,01	0,0000
		2	0,77	0,81	0,48	0,0520
		3	0,77	0,81	0,50	0,0540
		4	0,77	0,81	0,53	0,0570
	0.33	1	0,77	0,77	0,01	0,0000
		2	0,77	0,83	0,62	0,0900
		3	0,77	0,83	0,64	0,0940
		4	0,77	0,83	0,66	0,0980
	0.50	1	0,77	0,77	0,02	0,0000
		2	0,77	0,88	0,91	0,2130
		3	0,77	0,87	0,93	0,2180
		4	0,77	0,87	0,94	0,2210

FIGURE 5

SIMULATION STUDY RESULTS



accuracy and discrimination of two regression methods. The former, GEE, is quasi-likelihood based in that it only requires specification of the mean and variance, not the entire probability distribution [7,10]. The latter, MTM, allows for simultaneous likelihood-based estimation of the average response and for the serial dependence among longitudinal observations. The latter model is permissible both when subjects have varying lengths of follow-up and when data may be missing at random [14], as it could be considered the case in the OM cohort. Instead, due to the presence of 'non-ignorable' drop outs due to deaths in CV data, a sensitivity analysis considering all different group of patients was performed, and no relevant differences in estimates and accuracy across groups between the two methods emerged. Estimates of the regression coefficients were similar for the two approaches, both on real case studies and in the MSE values derived from the simulation. However, evident differences were observed in discrimination, determined by the better characterization at the individual-level of a serial dependence that compared a person's potential outcomes under different paths of interest in MTM. Calibration at large was globally similar between the two models, particularly suffering in the CV data for the sub-optimal linear modelling of the time effect, and showing in both case studies a higher precision of the GEE approach at the end of the longitudinal path. No significant differences in calibration between the two methods emerged from the simulation study.

Results of the present study confirm that the practical choice of the regression approach should depend on the scientific question being addressed: in CV data the primary goal was in the individual estimation of the right timing of ICD implantation; a "retrospective" scenario was considered in which longitudinal data were available: by looking at transitions between couple of time points an important improvement in the first 6-12 months and then a stabilization during the subsequent period was observed, even though a little, clinically negligible, worsening towards the 4th follow-up year was present, possibly due to the natural progression of the disease. The important role of the initial condition together with a not ignorable serial dependence across repeated measurements both emerged as strong individual predictors of longitudinal evolution. In this case, the MTM approach should be preferred with respect to GEE. For OM data, more than subject-specific, the cross-sectional population-average estimates lend themselves to an important interpretation: regulatory authorities would generally be mainly interested in determining the boundary values for the internal lumbar load to prevent the occurrence of sciatic pain that must show efficacy for the average population under study, and in this respect the GEE approach could be safely used.

Last but not least, the three measures used to evaluate model's performance, the Area Under the Curve (AUC), the Integrated Discrimination Improvement (IDI) and the Net Reclassification Improvement (NRI), in its continuous version, offer complementary information to evaluate discrimination and have been here extended, for the first time at our knowledge, to the case of longitudinal binary regression models.

References

[1] Diggle PJ, Heagerty P, Liang KY, Zeger SL (2002). Analysis of Longitudinal Data. 2nd ed. Oxford: Oxford University Press.

[2] Merlo M, Pyxaras SA, Pinamonti B, Barbati G, Di Lenarda A, Sinagra G. Prevalence and prognostic significance of left ventricular reverse remodeling in dilated cardiomyopathy receiving tailored medical treatment. J Am Coll Cardiol. 2011 Mar 29;57(13):1468-76.

[3] Bovenzi M, Metrics of whole-body vibration and exposure-response relationship for low back pain in professional drivers: a prospective cohort study. Int Arch Occup Environ Health 2009, 82:893-917.

[4] Bardy, GH. The Sudden Cardiac Death–Heart Failure Trial (SCD-HeFT). In: Arrhythmia Treatment and Therapy: Evaluation of Clinical Trial Evidence. Eds: Woosley, RL and Singh, SN. 2000, 323-42.

[5] Hinz B, Seidel H, Blüthner R, Menzel G, Hofmann J, Gericke L,Schust M. Whole-body vibration experimental work and biodynamic modelling. Annex 18 to VIBRISKS Final Technical Report: Risks of Occupational Vibration Exposures - VIBRISKS. FP5 Project No. QLK4-2002-02650. European Commission Quality of Life and Management of Living Resources Programme 2007. Available from: http://www.vibrisks.soton.ac.uk.

[6] VIBRISKS, Risks of Occupational Vibration Exposures. FP5 Project No. QLK4-2002-02650. European Commission Quality of Life and Management of Living Resources Programme 2007. Available from: http://www.vibrisks.soton.ac.uk.

[7] Liang KY and Zeger SL, Longitudinal data analysis using generalized linear models. Biometrika, 1986, 73, 13-22.

[8] Heagerty PJ, Marginalized Transitions Models and Likelihood Inference for Longitudinal Categorical Data. Biometrics 2002, 58:342-51.

[9] Bovenzi M, A longitudinal study of low back pain and daily vibration exposure in professional drivers. Ind Health, 2010, 48:584-95.

[10] Zeger SL, Liang KY, Albert SP, Models for Longitudinal Data: A Generalized Estimating Equation Approach. Biometrics, Vol. 44, No. 4. (Dec., 1988), pp. 1049-60.

[11] [Azzalini A, Logistic regression for autocorrelated data with application to repeated measures, Biometrika, 1994, (81): 767-75.

[12] Yan J, Geepack: Yet Another Package for Generalized Estimating Equations. R News, 2002, 2, 12-4.

[13] Hu FB, Goldberg J, Hedeker D, Flay BF, Pentz MA, Comparison of Population-Averaged and Subject-Specific Approaches for Analyzing Repeated Binary Outcomes. American Journal of Epidemiology, 1998, Vol. 147, No. 7, pp 694-703.

[14] B. F. Kurland and P.J. Heagerty, Marginalized transition models for longitudinal binary data with ignorable and non-ignorable drop-out. Statist. Med. 2004; 23:2673–95.

[15] Hanley, James A.; McNeil, Barbara J., The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology, 1982, 143 (1): 29–36.

[16] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837-45.

[17] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS, Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond, Statist. Med. 2008; 27:157–72.

[18] Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P, Interpreting Incremental Value of Markers Added to Risk Prediction Models. Am J Epidemiol. 2012 Sep 15;176(6):473-81.

[19] Guerra MW, Shults J. A Note on the Simulation of Overdispersed Random Variables with Specified Marginal Means and Product Correlations. The American Statistician, Published online: 18 Feb 2014. DOI: 10.1080/00031305.2014.887592.