Longitudinal profile of a set of biomarkers in predicting Covid-19 mortality using joint models

Matteo Di Maso, Monica Ferraroni, Pasquale Ferrante, Serena Delbue, Federico Ambrogi

1. Introduction

In survival analysis, time-varying covariates (i.e., covariates that are repeatedly measured over time) are endogenous when (i) their measurements are directly related to the event status and (ii) when incomplete information occur at random points during the follow-up because subjects may skip schedule visits and dropout from the study (Rizopoulos, 2012). Consequently, the classical time-dependent Cox model (Therneau and Grambsch, 2000) leads to biased estimates.

In order to correctly estimate the association between a time-to-event outcome and endogenous covariates, two approaches become in widespread use. The first is the joint model (JM) for the simultaneously analysis of longitudinal and time-to-event data (Rizopoulos, 2012). In this approach, the survival sub-model (used to predict hazards for a set of time-invariant covariates) and longitudinal sub-model (used to predict time-varying covariates) are interdependent by means of a set of random effects (i.e., shared parameters). Random effects are individual-specific model terms, and their inclusion in JM provides a way of producing overall predictions. The second approach is the landmarking analysis (van Houwelingen and Putter, 2012), a more pragmatic method that avoids modelling the time-varying covariates. In this approach, the estimated effect of the time-varying covariates is based on the value at the landmark time point, after which values of time-varying covariates may change.

During the first wave of Covid-19 pandemic, physicians at Istituto Clinico di Città Studi in Milan collected a set of inflammatory biomarkers in order to understand what might be used as prognostic factors in progression and mortality of Covid-19 disease. Biomarkers were collected repeatedly over the follow-up. Furthermore, particularly in the first epidemic outbreak, physicians did not have standard clinical protocols for management of Covid-19 disease and for this reason, measurements of biomarkers were highly incomplete especially at the baseline.

The aim of the present study is twice. Using data on Covid-19 patients, we firstly evaluate the association of a single biomarker on Covid-19 mortality using JM, landmarking, and time-dependent Cox model in order to compare estimates. Second, we present JM estimates for the whole set of biomarkers collected on Covid-19 patients to evaluate their association on mortality risk.

Matteo Di Maso, University of Milan, Italy, matteo.dimaso@unimi.it, 0000-0002-6481-990X
Monica Ferraroni, University of Milan, Italy, monica.ferraroni@unimi.it, 0000-0002-4542-4996
Pasquale Ferrante, University of Milan, Italy, pasquale.ferrante@unimi.it
Serena Delbue, University of Milan, Italy, serena.delbue@unimi.it, 0000-0002-3199-9369
Federico Ambrogi, University of Milan, Italy, federico.ambrogi@unimi.it, 0000-0001-9358-011X

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Matteo Di Maso, Monica Ferraroni, Pasquale Ferrante, Serena Delbue, Federico Ambrogi, *Longitudinal profile of a set of biomarkers in predicting Covid-19 mortality using joint models*, pp. 191-196, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-461-8.36, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci (edited by), *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the on-site conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-461-8 (PDF), DOI 10.36253/978-88-5518-461-8

2. Methods

Theoretical framework of JM

According to the shared parameter approach, the JM consists of two sub-models: one to model the time-to-event outcome (survival sub-model) and the other to model the time-varying covariates (longitudinal sub-model).

The survival sub-model is a typical semi-parametric (or parametric) model for time-to-event outcome. Let T_i^* be the true event time for the i_{th} subject (with $i = 1,\ldots,N$), T_i be the observed event time, defined as the minimum of the potential right-censoring time C_i and T_i^*, i.e., $T_i = \min\left(T_i^*, C_i\right)$, and let $\delta_i = I\left(T_i^* \le C_i\right)$ be the event indicator. Furthermore, let $m_i(t)$ be the true and unobserved value of a single time-varying covariate at time t. The (proportional) hazards model is:

$$h_i\left(t^*\right) = h_0\left(t^*\right) \cdot \exp\left\{\beta_j X_{ij} + \alpha m_i(t)\right\}$$

where $h_0(\cdot)$ denotes the baseline risk function, X_{ij} is the set of j time-invariant covariates measured at baseline for the i_{th} subject, β_j is the corresponding vector of regression coefficients, and α is the regression coefficient for the time-varying covariate, quantifying the effect of such variable to the event risk.

The longitudinal sub-model is a typical linear mixed model for longitudinal outcome. As information on the time-varying covariate are collected intermittently and with error at a set of few time points for each subject, the aim of longitudinal sub-model is to predict the complete longitudinal history (also called trajectory) of the time-varying covariate (the outcome of the longitudinal sub-model) for a set of time-invariant covariates. In particular, longitudinal sub-model is:

$$y_i(t) = m_i(t) + \varepsilon_i(t)$$

where $m_i(t) = \gamma_j X_{ij} + g_i Z_i(t)$, with $g_i \sim N(0, D)$ and $\varepsilon_i(t) \sim N\left(0, \sigma^2\right)$. The quantity $y_i(t)$ is the observed longitudinal outcome for the i_{th} subject at time t, γ_j denote the estimates for the fixed effects X_{ij} and g_i denote the estimates for the random effects $Z_i(t)$. In the shared parameter approach, the random effects are common for the longitudinal and survival sub-models.

Recently, a Bayesian approach for fitting JM was introduced. In particular, estimation of JM's parameters proceeds using Markov chain Monte Carlo (MCMC) algorithm. The posterior distribution of the model parameters is derived under the assumptions that given the shared parameter, both longitudinal and survival sub-models are assumed independent, and the longitudinal outcomes of each subject are assumed independent. In this approach, non-informative priors can be used for explorative purposes.

Landmarking analysis

The idea behind landmarking analysis is to select, for a given time point $t_{LM} = s$, all subjects alive and under follow-up at time s. In particular, landmarking involves to set s and using the value of the time-varying covariate at s as fixed covariate in a time-dependent Cox model from s onwards, in a subset of subjects at risk at s. For a generic subject i, the objective is to use part of the information of the time-varying covariate of the subject to estimate the conditional probability that the subject is still alive after a predefined time window w. More specifically, at a prediction time point s, the conditional probability that the subject is still alive at time $w+s$ conditionally on being alive at time s and conditional the history of the time-varying covariate up to s is given by:

$$\pi_i\left(s+w \mid s\right) = P\left\{T_i > s+w \mid T_i \geq s, m_i\left(s\right)\right\}$$

with $m_i\left(s\right)$ denoting the history of time-varying covariate up to s.

Data collection

Between 21 February and 19 March 2020, a total of 403 Covid-19 patients were admitted at Istituto Clinico Città Studi in Milan. Patients aged 21-100 years and 58.3% were men. Person-time at risk was computed as the time elapsed from the day of hospital admission to the day of Covid-19 death (event time), to the day of hospital discharge, or to the day of moving in other structure (right-censoring time), whichever came first. Baseline characteristics included sex and age of patients, whereas biomarkers measurements included ferritin (ng/ml), lymphocytes count, neutrophil granulocytes count, D-dimer (ng/ml), C-reactive protein (ml/l), glucose (mg/dl) and lactate dehydrogenase (LDH; U/l).

Statistical analysis

In order to compare JM, landmarking, and time-dependent Cox model estimates, ferritin was considered. In particular, logarithm of ferritin (log-ferritin) levels was used to account for the skewedness of the measurements. According to the Bayesian approach, independent and non-informative priors for the fixed effects of the longitudinal and survival sub-models (i.e., age and sex) and for the shared parameter (i.e. subject-specific predicted trajectories of log-ferritin level) were used in the JM. In addition, a natural cubic spline with 2 knots was used to model the subject specific log-ferritin trajectories through time and to model age. Two knots are generally sufficient to detect mild non-linear effects and to avoid over-parametrization of the model considering the available sample size.

In landmarking analysis, a set of landmarking time point for log-ferritin time of measurements was considered. In particular, data were analysed with s running from 3 to 20 days which corresponded to the median and the 75th centile of log-ferritin time of the first and last measurements, respectively. Prediction windows w were set at 7, 14, 21, and 28 days. Age was modelled in the same way as the JM.

In the time-dependent Cox model, observed log-ferritin levels was incorporated as time-varying covariate using a natural cubic spline with 2 knots as well as age.

In order to provide associations between biomarkers and Covid-19 mortality, a JM of each biomarker one at time with the occurrence of Covid-19 death was performed (univariable JM). Logarithmic transformation was considered for ferritin, lymphocytes (log-lymphocytes), neutrophil granulocytes (log-neutrophil granulocytes), D-dimer (log-D-dimer), and C-reactive protein (log-C-reactive protein). In the multivariable JM including all biomarkers (multivariable JM), D-dimer was excluded due to the high number (78; 19%) of patients with missing values. Assumptions for priors, biomarkers trajectories and age were the same as the JM for log-ferritin previously described.

Analyses were performed using JMbayes (Rizopoulos, 2016) and dynpred (Putter, 2015) packages in R Statistical Software, version 4.0.5 (R Core Team 2021).

3. Results

Among 403 Covid-19 patients admitted at Istituto Clinico Città Studi, 140 patients died during the follow-up. Among 263 patients survived, 99 were discharged and 164 were moved in other structures. The median of follow-up was 14 days (range: 0-78 days).

Hazard ratios (HR) and corresponding 95% confidence intervals (CI) from the (biased) time-dependent Cox model and JM for log-ferritin levels (ng/ml) were 2.10 (1.67-2.64) and 1.73 (1.38-2.20), respectively. According to landmarking analysis, the HR was 1.73 (1.25-2.38) for a prediction window of 7 days. With regards to 14, 21, and 28 prediction windows, HRs were 1.86 (1.36-2.54), 1.91 (1.40-2.60), and 1.91 (1.40-2.61), respectively.

The estimates obtained from univariable JM showed decreased level through time for expected log-ferritin according to the negative coefficients for the splines of time at measurements (table 1). Conversely, the expected level of log-ferritin increased with increasing age and men showed higher expected levels than women. The expected log-lymphocytes count increased through time, whereas it decreased with age. No association emerged between log-lymphocytes and sex. The expected log-neutrophil granulocytes count decreased through time, whereas it increased with age and men showed higher levels. Likewise, expected log-D-dimer levels decreased through time, increased with age and men had higher levels. For log-C-reactive protein, expected levels showed a mixed trend through time. In particular, levels initially decreased according to the negative coefficient for the first part of follow-up and increased thereafter. The expected log-C-reactive protein levels increased with age and men showed higher levels than women. Expected levels of glucose and LDH decreased through time, while increasing with age and men had higher levels.

In univariable JM, all biomarkers were significantly associated with Covid-19 mortality. An increase in the levels of biomarkers was associated with an increased in the mortality risk, except for lymphocytes. In particular, doubling of levels for log-lymphocytes count was associated with approximately halving mortality risk (HR=0.58; 95% CI: 0.46-0.73). The strongest associations were observed for log-neutrophil granulocytes (HR=2.87; 95% CI: 2.30-3.51 for doubling of levels), for log-C-reactive protein (HR=2.44; 95% CI: 2.01-2.97 for doubling of levels) and glucose (HR=2.89; 95% CI: 1.92-4.26 for an increase of 100 mg/dl).

The multivariable JM was estimated using data on 320 patients with 96 (30%) events (after exclusion of patients with missing values for D-dimer). For ferritin and lymphocytes there were no more evidence of association with mortality. The strength of the association was attenuated with respect to the univariable JM for log-neutrophil granulocytes (HR=1.78; 95% CI: 1.16-2.69 for doubling of levels), log-C-reactive protein (HR=1.44; 95% CI: 1.13-1.83 for doubling of levels), LDH (HR=1.28; 95% CI: 1.09-1.49 for an increase of 100 UI/l), and glucose (HR of 2.44; 95% CI: 1.28-4.26 for an increase of 100 mg/dl).

However, the strongest effect in both univariable and multivariable JM was observed for age with a HR starting to rapidly increase approximately at 60 years.

4. Conclusion

In the present work, we firstly compared HR estimates of a single time-varying covariate (log-ferritin) using different approaches. The HRs from JM and landmarking approaches were lower than that of the time-dependent Cox model. In addition, landmarking estimate for a 7-day prediction window was similar to the estimate of the JM, but it tended to increase increasing prediction window. However, landmarking estimates were lower than the time-dependent Cox model one.

Finally, the multivariable JM model showed associations between some biomarkers and Covid-19 mortality but the strong association between age and mortality risk persisted after adjusted for biomarkers considered.

References

Rizopoulos D. (2012). *Joint Models for Longitudinal and Time-to-Event Data. With Application in R.* Boca Raton: Chapman & Hall/CRC.

Therneau T., Grambsch P. (2000). *Modeling Survival Data: Extending the Cox Model.* Springer-Verlag, New York (NY).

van Houwelingen HC., Putter H. (2012). *Dynamic Prediction in Clinical Survival Analysis.* Boca Raton: Chapman & Hall/CRC.

Rizopoulos D. (2016). The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data using MCMC. *J Stat Softw*. **72**(7), pp. 1-45.

Putter H. (2015). *dynpred: Companion Package to "Dynamic Prediction in Clinical Survival Analysis"*. R package version 0.1.2. https://CRAN.R-project.org/package=dynpred.

Table 1. Univariable and multivariable joint model estimates.

Variables	Univariable model		Multivariable model	
	Effect (95% CI)	**p-value**	**Effect (95% CI)**	**p-value**
Longitudinal process: log-ferritin (ng/ml)				
Intercept	5.18 (4.67, 5.67)	p<0.01	4.88 (4.35, 5.43)	p<0.01
ns(time in days, 2)1	-1.05 (-1.32, -0.77)	p<0.01	-1.02 (-1.70, -0.40)	p<0.01
ns(time in days, 2)2	-1.69 (-2.23, -1.08)	p<0.01	-1.93 (-3.55, -0.54)	p=0.01
Sex (male vs female)	0.53 (0.34, 0.70)	p<0.01	0.66 (0.47, 0.85)	p<0.01
ns(age in years, 2)1	2.13 (1.15, 3.06)	p<0.01	1.70 (0.70, 2.72)	p<0.01
ns(age in years, 2)2	0.20 (-0.18, 0.59)	p=0.31	0.02 (-0.39, 0.44)	p=0.95

Longitudinal process: log-lymphocytes count				
Intercept	0.68 (0.29, 1.05)	p<0.01	0.69 (0.42, 0.95)	p<0.01
ns(time in days, 2)1	1.11 (0.95, 1.27)	p<0.01	1.21 (0.88, 1.56)	p<0.01
ns(time in days, 2)2	0.52 (0.40, 0.65)	p=0.01	0.87 (0.35, 1.42)	p<0.01
Sex (male vs female)	-0.11 (-0.25, 0.03)	p=0.11	-0.15 (-0.25, -0.06)	p<0.01
ns(age in years, 2)1	-1.47 (-2.20, -0.72)	p<0.01	-1.25 (-1.74, -0.76)	p<0.01
ns(age in years, 2)2	-0.69 (-0.97, -0.39)	p<0.01	-0.51 (-0.72, -0.28)	p<0.01
Longitudinal process: log-D-dimer (ng/ml)				
Intercept	4.34 (3.62, 5.01)	p<0.01	-	-
ns(time in days, 2)1	-1.72 (-2.02, -1.39)	p<0.01	-	-
ns(time in days, 2)2	-2.99 (-3.47, -2.57)	p<0.01	-	-
Sex (male vs female)	0.35 (0.10, 0.61)	p=0.01	-	-
ns(age in years, 2)1	4.18 (2.93, 5.52)	p<0.01	-	-
ns(age in years, 2)2	1.57 (1.01, 2.09)	p<0.01	-	-
Longitudinal process: log-neutrophil granulocytes count				
Intercept	0.90 (0.59, 1.22)	p<0.01	0.86 (0.56, 1.15)	p<0.01
ns(time in days, 2)1	-1.10 (-1.41, -0.80)	p<0.01	-0.55 (-1.08, 0.05)	p=0.07
ns(time in days, 2)2	-2.88 (-3.50, -2.25)	p<0.01	-1.28 (-2.31, -0.01)	p=0.05
Sex (male vs female)	0.25 (0.14, 0.38)	p<0.01	0.29 (0.18, 0.40)	p<0.01
ns(age in years, 2)1	1.25 (0.63, 1.86)	p<0.01	0.84 (0.30, 1.39)	p<0.01
ns(age in years, 2)2	0.90 (0.64, 1.14)	p<0.01	0.55 (0.33, 0.78)	p<0.01
Longitudinal process: log-C-reactive protein (ml/l)				
Intercept	-0.18 (-0.70, 0.34)	p=0.50	-0.12 (-0.70, 0.43)	p=0.69
ns(time in days, 2)1	-4.43 (-5.30, -3.56)	p<0.01	-4.10 (-5.59, -2.52)	p<0.01
ns(time in days, 2)2	0.54 (-1.48, 2.58)	p=0.59	2.33 (-1.04, 6.10)	p=0.18
Sex (male vs female)	0.49 (0.31, 0.69)	p<0.01	0.49 (0.28, 0.70)	p<0.01
ns(age in years, 2)1	4.27 (3.23, 5.29)	p<0.01	3.45 (2.42, 4.52)	p<0.01
ns(age in years, 2)2	1.01 (0.60, 1.41)	p<0.01	0.52 (0.02, 1.00)	p=0.04
Longitudinal process: glucose/100 (mg/dl/100)				
Intercept	86.58 (83.00, 90.12)	p<0.01	77.27 (65.91, 88.76)	p<0.01
ns(time in days, 2)1	-19.15 (-27.92, -10.52)	p<0.01	-0.80 (-16.66, 14.61)	p=0.92
ns(time in days, 2)2	-10.09 (-20.68, 0.29)	p=0.06	-1.54 (-20.02, -17.47)	p=0.86
Sex (male vs female)	1.01 (0.95, 1.08)	p<0.01	12.33 (4.57, 19.78)	p<0.01
ns(age in years, 2)1	58.28 (57.94, 58.65)	p<0.01	36.10 (19.21, 53.01)	p<0.01
ns(age in years, 2)2	12.27 (12.13, 12.41)	p<0.01	2.82 (-10.77, 15.95)	p=0.67
Longitudinal process: LDH/100 (UI/100)				
Intercept	106.22 (1.40, 2.77)	p<0.01	97.42 (77.40, 116.33)	p<0.01
ns(time in days, 2)1	-26.77 (-43.74, -9.79)	p<0.01	8.24 (-10.46, 26.40)	p=0.36
ns(time in days, 2)2	-3.40 (-22.74, 16.14)	p=0.72	-7.02 (-27.95, 13.12)	p=0.49
Sex (male vs female)	23.52 (23.29, 23.73)	p<0.01	58.08 (40.70, 75.22)	p<0.01
ns(age in years, 2)1	243.14 (241.65, 243.90)	p<0.01	48.13 (27.32, 70.79)	p<0.01
ns(age in years, 2)2	81.11 (80.73, 81.49)	p<0.01	0.59 (-19.33, 19.95)	p=0.96

Variables	**log-hazard (95% CI)**	**p-value**	**log-hazard (95% CI)**	**p-value**
Time-to-event process				
Sex (male vs female)	-		0.74 (0.26, 1.20)	P<0.01
ns(age in years, 2)1	-		9.31 (3.35, 15.75)	p<0.01
ns(age in years, 2)2	-		3.82 (2.57, 5.26)	p<0.01
log-ferritin (ng/ml)	0.55 (0.33, 0.79)	p<0.01	-0.13 (-0.47, 0.22)	p=0.48
log-lymphocytes	-0.78 (-1.11, -0.44)	p<0.01	0.04 (-0.43, 0.53)	p=0.89
log-neutrophil granulocytes	1.52 (1.20, 1.81)	p<0.01	0.83 (0.21, 1.43)	p=0.01
log-C-reactive protein (ml/l)	1.29 (1.01, 1.57)	p<0.01	0.53 (0.18, 0.87)	p<0.01
glucose/100 (mg/dl/100)	1.06 (0.65, 1.45)	p<0.01	0.89 (0.25, 1.45)	p=0.01
LDH/100 (UI/l/100)	0.55 (0.46, 0.64)	p<0.01	0.25 (0.09, 0.40)	p<0.01