Modelling time-to-dropout via Shared Frailty Cox Models. A trade-off between accurate and early predictions

Masci, C.; Cannistrà, M.; Mussida, P.

Modelling time-to-dropout via Shared Frailty Cox Models. A trade-off between accurate and early predictions

Chiara Masci[a]; Marta Cannistrà[b*]; Paola Mussida[c]

[a] *MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano; Milan, Italy*

[b] *School of Management, Politecnico di Milano; Milan, Italy*

[c] *Department of Electronics, Information and Bioengineering, Politecnico di Milano - Milan, Italy*

[*] Corresponding author: marta.cannistra@polimi.it

Abstract. This paper investigates the student dropout phenomenon in a technical Italian university in a time-to-event perspective. Shared frailty Cox time-dependent models are applied to analyse the careers of students enrolled in different engineering programs with the aim of identifying the determinants of student dropout through time, to predict the time to dropout as soon as possible and to observe how the dropout phenomenon varies across time and degree programs. The innovative contributions of this work are methodological and managerial. First, the adoption of shared frailty Cox models with time-varying covariates is relatively new to the student dropout literature and it allows to take account of the student career evolution and of the heterogeneity across degree programs. Second, understanding the dropout pattern over time and identifying the earliest moment for obtaining its accurate prediction allow policy makers to set timely interventions for students at risk of dropout.

Keywords: Shared frailty Cox model; student dropout; survival analysis; early warning system; time to dropout; time-varying covariates.

Subject classification codes: I23

1. Introduction and Motivation

The Italian Higher Education system is affected by a high level of dropout, with many students abandoning their Bachelor programs during the first or second year. According to ANVUR (Italian National Agency for the Evaluation of Universities and Research Institutes), the Italian dropout rate for the students from whom complete data is available is around 24%, with half (12%) of them dropping out in the first two years (ANVUR, 2018).

This data is even more worrying considering that only 28% graduates in Italy are from the 25-34 years old population, against a European average of 40 University dropout represents a worrisome phenomenon with both economic and social impacts.

From the economic standpoint, dropout represents a net waste of resources for universities, since education is a costly activity. From the social perspective, dropout affects students, who face a social stigma (e.g., fewer job opportunities and lower salaries), disconnecting them with their social environment (Alban & Mauricio, 2019).

Hence, studying the dropout phenomenon and its determinants is paramount. Identifying students at risk and, in particular, the riskiest moment of their career is extremely important: only with timely interventions, universities would be able to retain their students, shepherding them towards graduation (Seidel & Kutieleh, 2017).

In line with the presented motivations, the aim of this paper is twofold: to study student's dropout and its major responsible factors across time and to identify the earliest moment in time in which we can provide dropout predictions that are (sufficiently) good both in terms of event and event time and can be used by universities for early intervention though appropriate preventive actions. The focus is on identifying not only who the students at risk are, but also when these students are at risk, discussing the effectiveness of an Early Warning System. The study is held at Politecnico di Milano (PoliMi) in Italy.

The paper is organized as follows: in Section 2, we set up an overview of the academic literature about survival analysis and dropout; in Section 3 we present the main features of the PoliMi dataset and the methodology adopted; results and final considerations are detailed in Sections 4 and 5.

2. Related Literature: survival analysis for studying dropout

As part of the wide academic literature aiming at predicting dropout in Education settings (Cannistrà et al., 2022; Hegde & Prageeth, 2018; Kehm, Larsen, & Sommersel, 2019), survival analysis is directed toward the deepening of when dropout occurs, considering the students' educational career complemented with its time dimension.

Many studies applying survival analysis focus on digital learning (Xie, 2019; Utami et al., 2020; Spitzer et al., 2021; Chenet al., 2020) for a simple reason: it is easy to track students over time, you know exactly when they drop out from the platform (i.e., last access). Hence, this method is less frequently applied to traditional settings. Another stream of literature is centered on the doctoral path: when and why PhD students drop out from their career (Van Der Haert et al., 2014; Booth & Satchell, 1995; De Valero, 2001; Grove, Dutkowsky, & Grodner, 2007). Indeed, the investigation of this phenomenon is relevant since it gives the opportunity to understand the most effective type of support for retaining PhD students, given their value in our society (Van Der Haert et al., 2014).

Focusing on schools and universities, the academic contributions applying survival analysis to students' academic career progression aim at modelling the phenomenon by highlighting the most important underlying factors (Arulampalam, Naylor, & Smith, 2004; Weybright et al., 2017; Thaithanan et al. , 2021; Patacsil, 2020; Min et al. 2011; Plank, DeLuca, & Estacion, 2008; Barragan, Gonza´lez, & Calderon, 2022; Vallejos & Steel, 2017; No, Taniguchi, & Hirakawa, 2016; Gury, 2011; Lesik, 2007).

Arulampalam et al. (2004) and Barragan et al. (2022) found academic performance to be an important dropout predictor, while according to Weybright et al. (2017), the student's background (e.g., being a male and not living with his mother) plays a significant role in predicting dropout (Barragan et al., 2022). Soares et al. (2015) observed that the difficulties faced with particular subjects, the desire for a different school, the perception that those completing their studies will have better job opportunities, and the importance assigned to school choice influence dropout from secondary school. When looking at the university dropout phenomenon's time component, Min et al. (2011) found significant differences for early semesters across groups. White and/or female students tend to leave university earlier than other sub-populations. Engineering students mostly abandon their academic career during the third semester, but this can happen even during the second semester when the student has a low math grade.

In terms of adopted models, the majority of scholars (Weybright et al., 2017; Thaithanan et al., 2021; Min et al., 2011; Plank et al., 2008; Barragan et al., 2022; Vallejos & Steel, 2017; Gury, 2011; Lesik, 2007; Arulampalam et al., 2004) use Cox PH models to estimate the probability of dropping out, often comparing Kaplan-Mayer curves on different students' features. Interesting sources of innovation are related to the comparison between fixed and random effects, as in Arulampalam et al. (2004), to model the effect of being enrolled in different degree programs; or to the combination between survival analysis and analytic hierarchy process methodologies, as in Barragan et al. (2022), to model dropout as a decision subjected to multiple alternatives; or by handling covariates' selection within a Bayesian framework (Vallejos & Steel, 2017). Generally, academic literature is moving toward modelling dropout and estimating its related factors with ever-increasing precision.

To contribute to this stream of research, this paper aims at studying the dropout phenomenon with a time perspective, adding two sources of innovation. The first relates the methods adopted, where the inclusion of frailties allows to account for the nested structure of students into degree programs, modelling the heterogeneity at the second level of the hierarchy, and the modelling of time-dependent covariates allows to update student information in time, building increasingly informed models. The second innovation regards the final collateral goal of the analysis: identifying the earliest moment in a student's career in which we can accurately predict his/her time-to-dropout. Indeed, early and accurate predictions are essential to effectively support at-risk students.

3. Data and Methods

In the following two subsections we present the dataset and our methodological approach.

3.1. PoliMi dataset

The PoliMi dataset contains administrative information about the careers of students enrolled between Academic Years 2010 and 2021 (12 years span period) in Bachelor's degree programs of Engineering. The University collects information about students' demographics and previous studies and tracks their entire academic careers, making anonymized data available in real time (Mussida & Lanzi, 2022). The demographics regard gender and age, residency and citizenship, and university's fee bracket paid by the student (as a proxy of socio-economic status). Then, high school track and final mark inform about student's previous career, while PoliMi admission test score is the first grade measured at the University. As regards career tracks, the number of credits obtained (ECTS) and the relative Grade Point Average (GPA) are collected for each student each semester.

The analysis excludes students who abandon their studies during the first semester of their first year, since many students enroll at PoliMi while waiting to be admitted to other programs at other universities, or they immediately decide to abandon because they had different expectations. This heterogeneity behind these dropouts might bias the results and these are not the dropouts that we aim to identify and on which we want to act[1].

[1] In the dataset, students who dropped out during their first semester are 1,700 (21.33% of the total dropout students). The University does not have time to take targeted preventive actions on these dropouts; therefore, their prediction is neither attractive nor valuable.

The final dataset contains 49,501 students enrolled within 16 degree programs. Table 1 reports the selected student-level variables, collected at the time of enrolment, with their explanation and summary statistics. The target variable regards the status of the student's career at the end of the third year, which can be concluded with graduation, with a dropout or with the student still being active. Variables Status at 3y and Career duration at 3y, reported in Table 1, define the target variable. Regarding the career tracks, Table 2 reports the selected longitudinal information relative to each student's careers, semesterly updated, that are ECTS and GPA.

The distribution of the students within the 16 Engineering degree programs is reported in Table A1a in the Appendix A1. For privacy reasons, we are not allowed to report the degree programs names but only their anonymized codes.

Table 1: Student-level variables adopted in the analysis, their description, type, and summary statistics.

Name	Description	Type	Summary info
Gender	Student gender (F/M)	Categorical	Male=77,5%, Female=22,5%

Admission age	Age as of the day of enrolment	Numerical	mean=18.72,
			median=19, sd=1.22,
			range=[16-61]

Income	University fee bracket: *High*, *Medium*, *Low* or *Study Grant (SG)*	Categorical	High=32.8%, Medium=23.5%, Low=30.6%, SG=13.1%

Origins	*Milanese* living in Milan, *Commuter* living outside Milan, *Offsite* have moved to Milan	Categorical	Commuter=67.5%,
			Milanese=25.7%,
			Offsite=6.8%

Highschool type	Field of study at high school: *Scientific*, *Classic*, *Technical*. *Foreigner* if he/she got his/her diploma abroad and *Other* if none of the above	Categorical	Classic=5.4%, Other=1.6%,
			Scientific=80.5%,
			Foreigner=0.7%,
			Technical=11.8%

Highschool grade	Grade obtained in high school	Numerical	Mean=84.87,
			median=85, sd=11.61,
			range=[60-100]

Admission score	Score obtained on the PoliMi admission test	Numerical	Mean=73.22,
			median=71.55, sd=9.36,
			range=[60-100]

Department	Study program of the student	Categorical	16 faculties

Status at 3y	Student career status considering a follow up time of 3 years, grouped by G (graduated), A (active), and D (dropout)	Categorical	Graduated=10.9%,
			Active=77.4%,
			Dropout=12.7%

Career duration at 3y	Length of the student career considering a follow up time of 3 years, expressed in semester	Numerical	Mean=5.08, median=6,
			sd= 1.47, range=[1,6]

Note: The Table shows the descriptive statistics of the time-invariant covariates used in subsequent analysis. In detail, for categorical variables it shows the distribution in each category, for Numerical variables their mean, median, standard deviation (sd) and range. Variables *Status at 3y* and *Career duration at 3y* are used to build the outcome of interest.

Table 2: Student-level variables related to the student career, measured each semester until the end of the third year.

	Description	Type	Summary info
Exa_Ay	Academic year corresponding to the observation	Categorical	Range=2010-2021,
Exa_Semester	Semester corresponding to the observation. 1 if first semester, 2 if second semester	Categorical	1=58.6%, 2=41.4%

ECTS	ECTS obtained by the student during each semester	Discrete	mean=18.48,
			median=20,
			range=0-40, sd = 12.6

GPA	Weighted average grade measured for each semester	Numeric	mean=18.97,
			median = 22.8,
			range=0-30, sd= 10.43

Note: the Table shows the structure of the time-dependent dataset, in which the GPA and ECTS are measured within each semester and Academic Year.

3.2.Models and Methods

In this subsection, we briefly recall the basics of survival analysis and we describe the statistical models adopted in the study.

3.2.1. Basics of survival analysis

Survival analysis regards the group of statistical procedures for the modelling of the time until an event of interest occurs (Kleinbaum & Klein, 1996; David & Mitchel, 2012)^{i}. For each unit of analysis, the event (i.e. *dropout*) might occur during the follow-up (i.e., the period of observation - in our case, three years) or not. In the second case, we refer to the observation as censored. For each unit i = 1,…, N, the target variable is defined as the couple of the survival time $T_i = \min(T_i\ast, C_i)$ and the censoring indicator $\delta_i = (T_i\ast \leq C_i)$, where C_i is the censoring time and T_i* is the observed event time, if any. δ_i is the indicator function that indicates whether the event occurred ($\delta_i = 1$) or not ($\delta_i = 0$) for the individual *i*. Censoring is assumed independent of survival time. Being T a non-negative random variable, the survival function

$$S(t) = \mathrm{P}(T > t) = 1 - \mathrm{P}(T \leq t) = 1 - F(t)$$

represents the probability of survival until time t, while the hazard function describes the instantaneous risk of failure and is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}.$$

The survival function S(t) can be estimated through the Kaplan-Meier estimator (KM) (Kaplan & Meier, 1958), which represents the probability of surviving in a given length of time while considering time in many small intervals. In case of two or more groups, the Log- Rank Ratio test (Mantel, 1966) can be used to test statistical differences across the estimated KM curves.

3.2.2. Shared frailty Cox PH models with time-invariant and time-varying covariates

Cox regression models are the most popular mathematical modelling approach to estimate the survival curves when considering several explanatory variables simultaneously. When the units are not *i.i.d.* but they are nested within groups, Shared Frailty Cox models introduce a frailty term, shared among units within the same group (in our case, students within degree programs), to take the structure into account (Kleinbaum & Klein, 1996; David & Mitchel, 2012).

The *Shared Frailty Cox Proportional Hazards (PH) model* assumes the hazard function for the i–th individual, for $i = 1,\ldots,N$ within the j–th group, for $j = 1,\ldots,J$, to be modelled as follows:

$$h_{ij}(t, \boldsymbol{x}_{ij}) \; = \; h_0(t) \times \omega_j \times exp\left\{\sum_{p=1}^{P} \beta_p x_{p,i,j}\right\}$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{\beta}\mathbf{x}_{ij}$ is the linear predictor, where $\mathbf{x}_i$ is the vector of the $i-$th individual P covariates and $\boldsymbol{\beta}$ is the vector of corresponding coefficients, ω_j is the frailty term for the $j-$th group. To better quantify the effect of the covariates, Hazard Ratios (HRs) can be derived from the vector of coefficients β. The modelling is based on the following assumptions: the effect of each covariate is constant across time (PH assumption), all failure times are independent given the frailties, and the values of the random effects ω_j are constant over time and common to all the individuals belonging to the same group. The frailties ω_j have a positive unobserved multiplicative effect on the hazard function. They are *i.i.d.* following a Gamma distribution with $E(\omega) = 1$ and $Var(\omega) = \theta$, where θ is the unknown parameter. Larger values of θ mean greater heterogeneity among the groups. Individuals belonging to a group with $\omega_j > 1$ have an increased hazard and decreased probability of survival compared to those with average frailty ($\omega_j = 1$). Similarly, individuals belonging to a group with $\omega_j < 1$ have a decreased hazard and increased probability of survival compared to those with average frailty.

This modelling can be extended to handle time-varying covariates. The shared frailty Cox model with both time-invariant and time-varying covariates, with respect to the *i*-th individual, assumes the following form:

$$h_{ij}\ (t, \boldsymbol{x_{ij}}(t)) = h_0(t) \times \omega j \times exp\left\{\sum_{p=1}^{P} \beta_p x_{p,i,j} \sum_{q=1}^{Q} \gamma_q x_{q,i,j}(t)\right\}$$

where P and Q are the number of time-invariant and time-varying covariates, respectively, and **β** and **γ** are the coefficients associated to these covariates, respectively. This modelling assumes that the effect of time-varying covariates $\boldsymbol{x}_q\ (t)$ on the survival probability at time *t* depends on the value of this feature at time *t* and not on its value at previous times. The PH assumption is no longer satisfied and the Hazard Ratio between two individuals *i* and *j* varies across time, depending on the covariates' values.

3.2.3. Goodness-of-fit indices

To evaluate the goodness of fit of our models, we rely on the most common metrics, the Concordance index (C-index) (Steck et al., 2007), which is defined as the proportion of concordant pairs, i.e., pairs of individuals for which the expected event times are predicted in the correct ordering, divided by the total number of possible evaluation pairs. The closer to one, the more accurate the Cox model. We support the C-index with a further evaluation, obtained by treating our survival models as classification models: by looking at the estimated survival probability at a fixed time t∗, we compute classification performance indices, e.g., precision, recall and ROC curve.

4. Results

The event of our interest is the *failure event of student dropout* from university. A follow up period of five semesters is considered: a student dropping out between the end of the first and sixth semester is labelled as *dropout*, while all other students, i.e., students who drop out after 3 years from the enrolment, who graduate or who have an active career at the end of the 3rd year, are marked as *censored.*

This section is divided into three main parts. In Section 4.1 we report results of a preliminary analysis to describe the cohort of students and the dropout distribution across time. In Sections 4.2.1 and 4.2.2 we show the results of shared frailty Cox models, first with only the time-invariant covariates and, then, with the addiction of the time-varying. Results focus on the interpretation of the effect of student-level characteristics on the dropout risk, on the quantification of the heterogeneity across degree programs and on the models' predictive power. Lastly, Section 4.3 reports a comparison of Cox models fitted by sequentially adding students information in time in order to identify the best trade-off between accurate and early predictions.

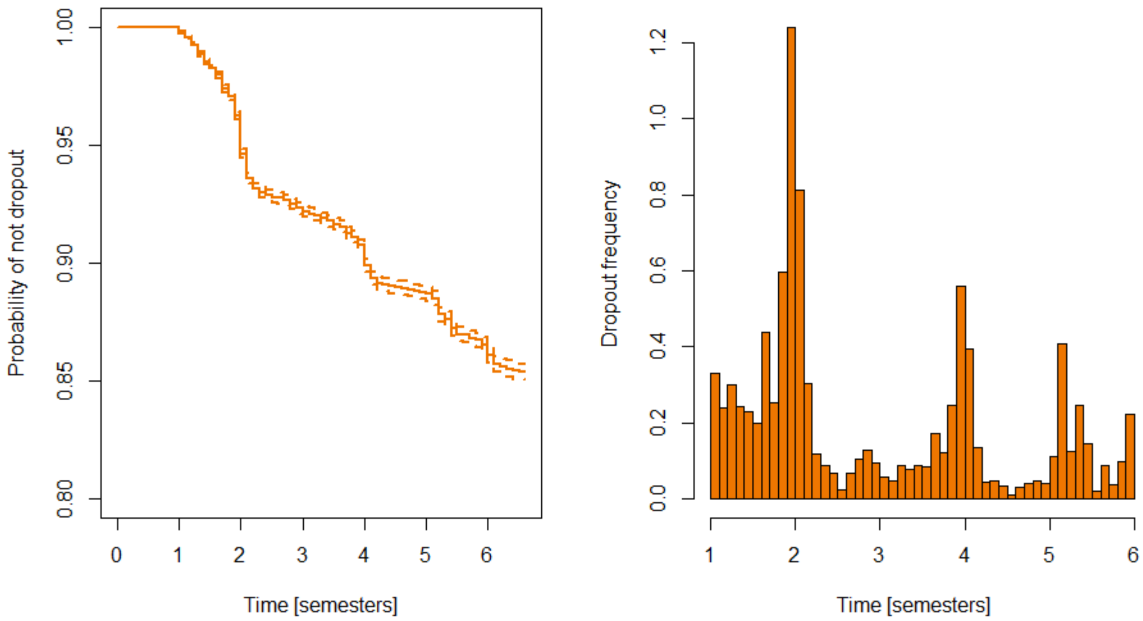
4.1.Preliminary analysis

As reported in Table 1, 12.7% of the students in our sample dropped out during the five semesters after the first one. Figure 1 reports the estimated survival function and the distribution of the time do dropout, measured in semesters. As expected, most of the dropouts occurs in correspondence of the end of academic years, mainly during the first two ones.

The hazard function presents three major peaks (represented by jumps in the survival function), which correspond to moments with high frequency of dropouts, at the end of each of the three academic years. The highest peak is in correspondence of the second semester, which marks the end of the first year of university; therefore, preventive interventions before this time are needed.

In order to investigate the association between student characteristics and dropout risk, we conducted a univariate analysis by computing KM survival curves. Figures in AppendixA2 report the KM curves computed for all student characteristics listed in Table 1, for the students' performance (ECTS and GPA) during the first semester and by Department. Figures show, for all students' features, different survival profiles.

Figure 1: Estimated survival function and time-to-dropout distribution.



Note: The figure in the right panel reports the distribution of the times, expressed in semesters, in which students definitely abandon PoliMi during a follow-up of 5 semesters (3 years, except for the first semester). Mean = 3.02, median = 2.20). 0 corresponds to the enrolment.

4.2.Shared frailty Cox PH models

In this section, we fit two Shared Frailty Cox models, considering students (level 1) nested within degree programs (level 2), in order to estimate the student time to dropout between the end of first and sixth semester, by exploring the effects of student characteristics and of the degree programs. The first is a Shared Frailty Cox model with time-invariant covariates, while in the second time-varying covariates about students' academic results are added.

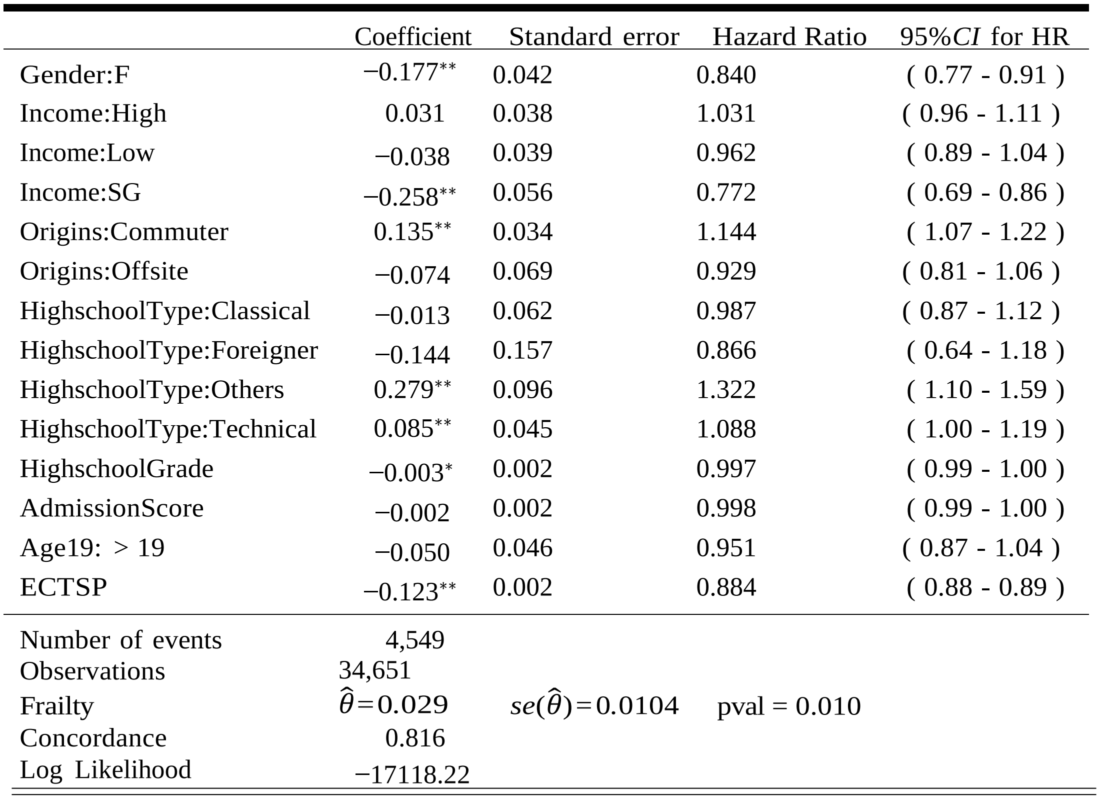
For both models, we randomly divide the dataset in training and test sets, containing 70% and 30% of the observations, respectively.

4.2.1. Shared Frailty Cox PH model with time-invariant covariates

The Shared Frailty Cox model includes as time-invariant covariates Gender, Income, Origins, HighschoolType, HighschoolGrade, AdmissionScore, Age19, and ECTS[2] of first semester. Table 3 shows the summary of the model estimated on the training set, where a total of 4,549 dropout events occurred. Results are in line with the ones of the KM curves. Females have an average lower risk of dropout than males (HR = 0.84), students with SG income category are less likely to drop out with respect to students in the Medium category (HR = 0.772), Commuters are more likely to drop out than Milanese students (HR = 1.144), being that a student who attended a Technical school or other types of high schools is associated to a higher dropout risk with respect to students who attended Scientific schools (HR = 1.088 and 1.322, respectively), and the higher the high school final grade, the lower the risk of drop out, on average (HR = 0.997). Lastly, the number of credits obtained at the first semester confirms to be an important protective factor. This output confirms again how the early academical results obtained by the student have an important role in a student's choice of withdraw from studies. The admission score at PoliMi and the age as of enrolment do not result to be significant. Figure A1 reports the *baseline survival and hazard functions* estimated by the model.

[2] We do not include *GPA* at this stage because *ECTS* and *GPA* are highly correlated, due to all students that have 0 *GPA* and 0 *ECTS* at first semester.

Table 3: Shared Frailty Cox model with time-invariant covariates, output of the summary



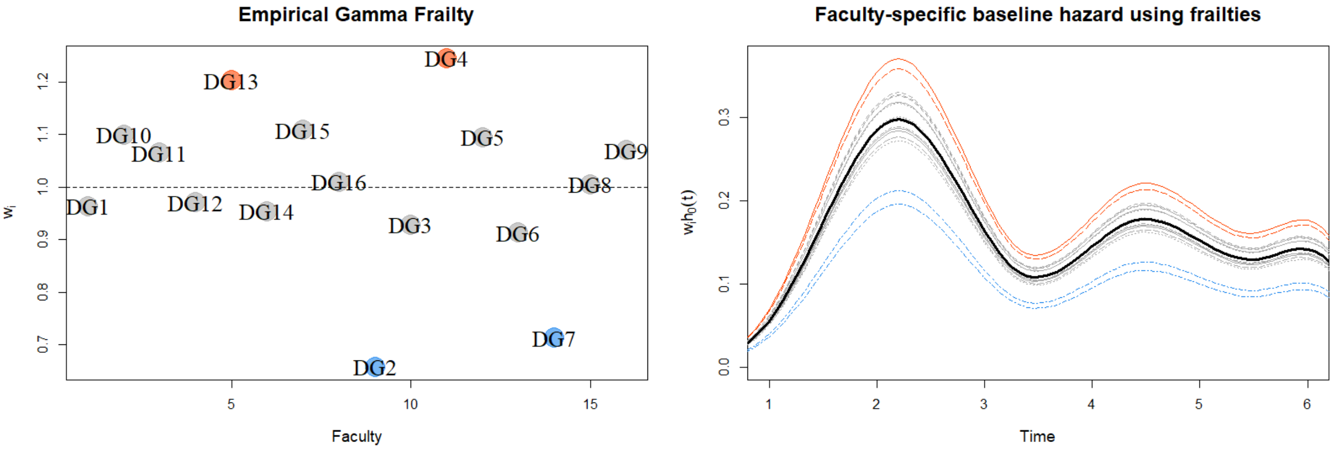
	Coefficient	Standard error	Hazard Ratio	95%*CI* for HR
Gender:F	-0.177^{**}	0.042	0.840	(0.77 - 0.91)
Income:High	0.031	0.038	1.031	(0.96 - 1.11)
Income:Low	-0.038	0.039	0.962	(0.89 - 1.04)
Income:SG	-0.258^{**}	0.056	0.772	(0.69 - 0.86)
Origins:Commuter	0.135^{**}	0.034	1.144	(1.07 - 1.22)
Origins:Offsite	-0.074	0.069	0.929	(0.81 - 1.06)
HighschoolType:Classical	-0.013	0.062	0.987	(0.87 - 1.12)
HighschoolType:Foreigner	-0.144	0.157	0.866	(0.64 - 1.18)
HighschoolType:Others	0.279^{**}	0.096	1.322	(1.10 - 1.59)
HighschoolType:Technical	0.085^{**}	0.045	1.088	(1.00 - 1.19)
HighschoolGrade	-0.003^{*}	0.002	0.997	(0.99 - 1.00)
AdmissionScore	-0.002	0.002	0.998	(0.99 - 1.00)
Age19: > 19	-0.050	0.046	0.951	(0.87 - 1.04)
ECTSP	-0.123^{**}	0.002	0.884	(0.88 - 0.89)
Number of events	4,549			
Observations	34,651			
Frailty	$\hat{\theta}=0.029$	$se(\hat{\theta})=0.0104$	pval = 0.010	
Concordance	0.816			
Log Likelihood	-17118.22			

Note: *p<0.1; **p<0.05. *Estimated baseline survival and hazard functions are reported in Figure A3a in Appendix A3.*

Regarding the degree program effect, $\hat{\theta}$ = 0.029 is the estimated variance of the frailty parameter. The variance of the frailty term $\hat{\theta}$ is significantly different from 0 (p-value of the Wald test 0.01), confirming the presence of heterogeneity between degree programs. The estimated frailty terms ω_j *, j = 1,…,16*, which denotes the effect of each particular study program on the baseline hazard function, are shown in left panel of Figure 4. Among the 16 degree programs, two result to be associated to a higher dropout risk with respect to the average, net to the effect of student characteristics (ω_{DG4} = 1.245 and ω_{DG13} = 1.203). On the opposite, two programs result to be associated to lower dropout risks ($\omega_{DG2} = 0.656$ and $\omega_{DG7} = 0.712$). In the plot, the groups are colored depending on the asymptotic 95% confidence interval [$\omega \pm 1.96 \times \sigma(\omega_j)$]. The groups whose lower bound of the confidence interval is greater than 1 are red, while the groups whose higher bound of the confidence interval is lower than 1 blue. In grey we find the departments whose confidence interval contains 1, suggesting that they are not significantly different from the average.

The impact of these estimated values on the survival probability can be easily visualized in the department-specific baseline hazard functions, showed in the right panel of Figure 4.

Figure 4: Estimated frailty terms and degree programs-specific baseline hazard functions in the time-invariant case



Note: Left panel shows the empirical Gamma Frailty terms for the 16 degree courses estimated by the shared frailty Cox model with time-invariant covariates. Red and blue points identify the faculties that have a frailty term significantly higher and lower than 1, respectively. Right panel reports the faculty-specific baseline hazard functions for the 16 specific degree courses.

In terms of model predictive performance, the C-index computed both on the training and test set are 0.816 and 0.814, respectively.

4.2.2. *Shared Frailty Cox PH model with time-varying covariates*

We now extend the previous model by including time-varying covariates. In particular, we consider GPA and ECTS measured at the end of each semester as time-progressive information. Model results are reported in Table 4.

By including the career tracks over time, some of the personal student characteristics change their significance with respect to the first model. Here, gender is no more significant; with respect to a Medium income, having a Low income and having a scholarship (SG) are protective factors; with respect to *Milanese* students, *Commuters* and *Offsite* students have on average a higher dropout risk; with respect to scientific high school, having attended a foreigner or a technical school is a protective factor; having obtained a good high school grade is a risk factor, and being a student older than the average is a protective factor. As regards the career track, both ECTS and GPA are very significant and are protective factors. It is worth to note that, net to the effect of progressive ECTS and GPA, we still observe many significant student characteristics.

Table 4: Shared Frailty Cox model with time-dependent covariates, output of the summary

	Coefficient	Standard Error	Hazard ratio	95%CI for HR
Gender:F	-0.06	0.043	0.945	(0.87-1.03)
Income:High	-0.053	0.038	0.948	(0.88-1.02)
Income:Low	-0.105^{**}	0.039	0.899	(0.83-0.97)
Income:SG	-0.293^{**}	0.056	0.746	(0.67-0.83)
Origins:Commuter	0.156^{**}	0.034	1.169	(1.09-1.25)
Origins:Offsite	-0.128^{*}	0.069	0.880	(0.77-1.01)
HighschoolType:Classical	0.028	0.062	1.029	(0.91-1.16)
HighschoolType:Foreigner	-0.337^{**}	0.157	0.713	(0.52-0.97)
HighschoolType:Others	0.144	0.096	1.155	(0.96-1.39)
HighschoolType:Technical	-0.144^{**}	0.045	0.866	(0.79-0.95)
HighschoolGrade	0.007^{**}	0.001	1.007	(1.00-1.01)
AdmissionScore	-0.003	0.002	0.997	(0.99-1.00)
Age19: > 19	-0.384^{**}	0.046	0.681	(0.62-0.75)
ECTSPprog	-0.061^{**}	0.001	0.941	(0.94-0.94)
GPAprog	-0.028^{**}	0.002	0.972	(0.97-0.98)
Number of events	4549			
Observations	197591			
Frailty	$\hat{\theta} = 0.012$	$se(\hat{\theta}) = 0.006$	pval = 0.020	
Concordance	0.857			
Log Likelihood	-14528.61			

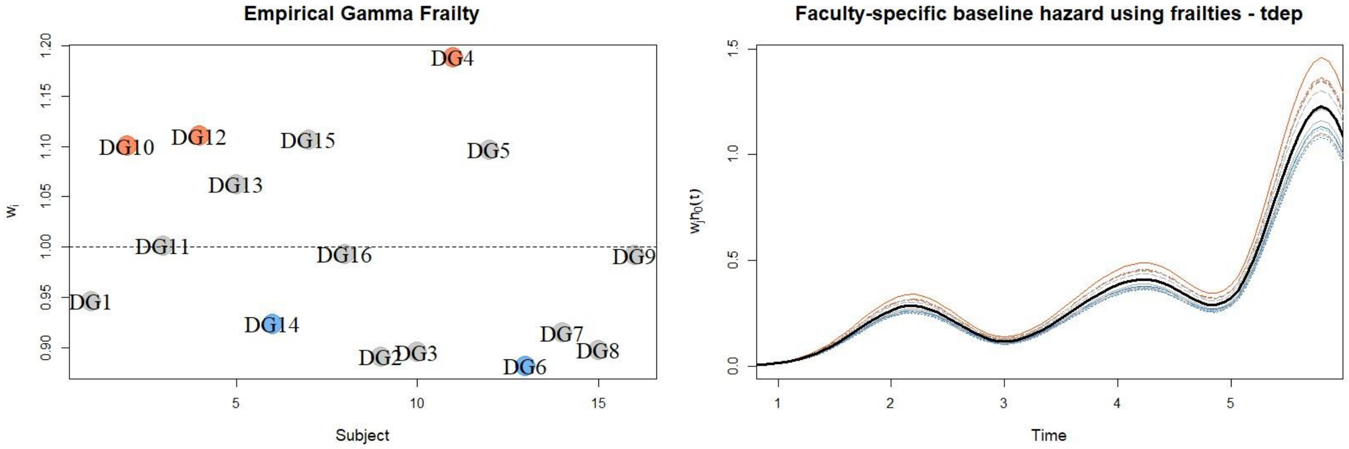
Note: *p<0.1; **p<0.05. *Estimated baseline survival and hazard functions are reported in Figure A3b in Appendix A3.*

The estimated *baseline survival and hazard functions* are displayed in Figure A2. At the end of the follow-up, the baseline survival function reaches very low value (and in parallel, the hazard reaches very high ones) with respect to the ones of the first model shown in Figure A1. The number of progressive ECTS mainly drives this trend since surviving until the end of the third year with 0 ECTS is very unlikely.

Regarding the frailty term, its estimated variance $\hat{\theta}$ = 0.012 again results to be significantly different from 0. The distribution of the 16 estimated frailties $\hat{\omega}_j$ and the program-specific baseline hazard functions are reported in Figure 5. Except for *DG*4, that confirms to be associated to a higher dropout risk both in the time-invariant and time-dependent frameworks, the other departments with an effect significantly different from 1 differ from the ones identified in the time-invariant framework. Here, *DG*10 and *DG*12 are associated to higher dropout risks, while *DG*6 and *DG*14 are associated to lower ones, suggesting that, net to the effect of the entire student career in the first three years, there are heterogeneous dropout dynamics across these departments.

The C-index measured on the training and on the test sets are both equal to 0.857. As expected, the inclusion of the career tracks over time improves the model accuracy and the predictive power, leading to a powerful model. Nonetheless, in order to promptly help at-risk students, early predictions are needed. In this perspective, in the next subsection we conduct a comparative analysis in order to estimate the best trade-off between accurate and early predictions.

Figure 5: Estimated frailty terms and faculty-specific baseline hazard functions in the time-varying case



Empirical Gamma Frailty

Faculty-specific baseline hazard using frailties - tdep

Note: Left panel shows the empirical Gamma Frailty terms for the 16 degree programs estimated by the shared frailty Cox model with time-varying covariates. Red and blue points identify the departments that have a frailty term significantly higher and lower than 1, respectively. Right panel reports the department-specific baseline hazard functions for the 16 specific degree programs.

4.3.Definition of an efficient Early Waning System

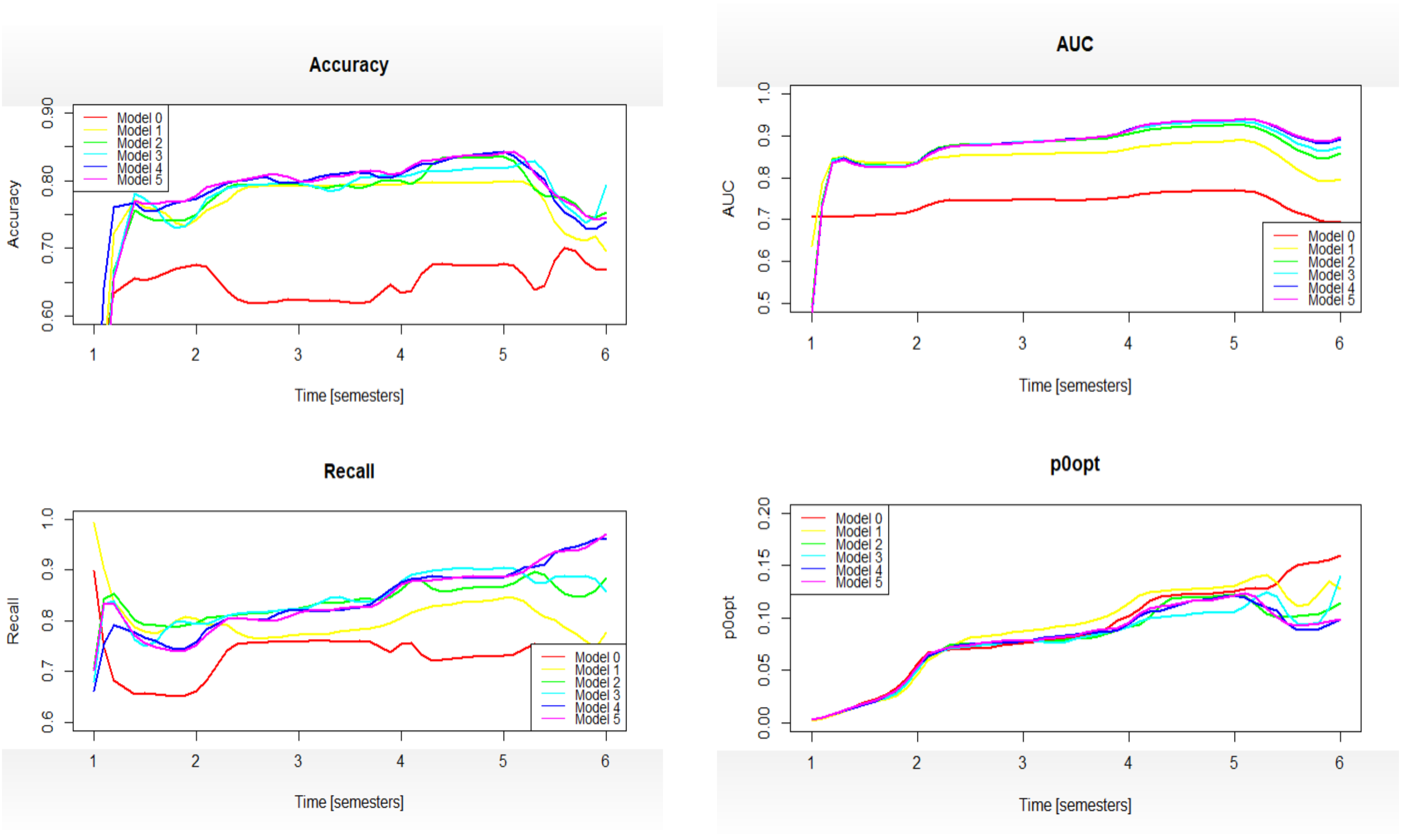
In order to evaluate the trade-off between early and accurate predictions, we perform a comparative analysis in which we build several shared frailty Cox models by including student information measured until different time points and we evaluate their predictive performance, in terms of C-index and classification indices. In particular, we build six subsequent models:

- *Model 0* only includes the student information measured at the time of enrolment (all time-invariant, no student career progress information is considered);

- *Model 1* includes the student information measured at the time of enrolment plus the number of ECTS obtained during the first semester (all time-invariant);
- *Model 2* includes the student information measured at the time of enrolment plus the progress of the number of ECTS and GPA obtained during the first two semesters;
- *Model 3* includes the student information measured at the time of enrolment plus the progress of the number of ECTS and GPA obtained during the first three semesters;
- *Model 4* includes the student information measured at the time of enrolment plus the progress of the number of ECTS and GPA obtained during the first four semesters;
- *Model 5* includes the student information measured at the time enrolment plus the progress of the number of ECTS and GPA obtained during the first five semesters;
- *Model 6* includes the student information measured at the time of enrolment plus the progress of the number of ECTS and GPA obtained during the first six semesters.

As we did in the previous section, we randomly divide the sample into training (70%) and test sets (30%). The predictive performance is measured in terms of C-index, accuracy, recall, and Area Under the ROC Curve (AUC), measured on the test set. For each of the six models, the classification indices are built at different time instants t* = {1.0, 1.1, 1.2,…, 5.9, 6.0} by classifying a student as dropout or not standing on his/her predicted dropout probability at time t*. At each time t*, the optimal threshold $p_0(t^*)$ for the classification is found (on the training set) and students in the test set are classified accordingly. Figure 6 shows the six trends of accuracy, recall, AUC, and optimal classification threshold in time, while Table 5 reports the C-index of the six models computed on the test set.

Figure 6: Estimated accuracy, recall, AUC, and optimal classification threshold in time for the six Shared frailty Cox models.



Note: Figures show Accuracy, Recall, AUC and values of optimal p for the predictions of dropout in different moments.

From Figure 6 and Table 5, we observe a first significant improvement in the models predictive performance when we move from Model 0 to Model 1 and a second less pronounced one when we move from Model 1 to Model 2. The difference in the performances between the last four models is instead almost negligible. This result suggests that student information at the time of enrolment is not sufficient to provide a good prediction for the dropout risk (Cindex = 0.682, accuracy between 0.65 and 0.7, AUC between 0.7 and 0.75). With the inclusion of first semester information, we become much more confident in identifying students at risk (Cindex= 0.813, accuracy between 0.7 and 0.75, AUC approximately 0.8), and with the entire first year information we reach a level that is comparable to the one that we obtain by observing the complete student career of the first six semesters.

This evidence, together with high frequency of dropout during the first year, suggest that first year career is already extremely informative and is enough to outline targeted interventions. The end of first and of the second semester represent two pivotal moments to implement preventive actions.

Table 5: Concordance Index computed on the test set, comparison between the 6 different time-dependent shared frailty Cox models.

Model	C-index
Model 0	0.682
Model 1	0.813
Model 2	0.849
Model 3	0.851
Model 4	0.855
Model 5	0.857

5. Concluding remarks and policy implications

The need to deal with the dropout issue is particularly relevant for scholars and policy makers, due to its important consequences at the personal, social, and economic levels (Castro-Lopez et al., & Bernardo, 2022). Early Warning System is a promising approach aiming at reducing educational withdrawal, predicting the phenomenon as soon as possible. However, academic literature focuses much on identifying the “who”, while less is done about the “when”. Indeed, the key research goals of this paper are identifying the time when dropout occurs and the optimal time to predict it.

To pursue these goals, we developed a set of shared frailty Cox models with time-invariant and time-varying covariates for predicting student dropout at different engineering faculties of PoliMi. The main innovation of this work relies on the methodological approach adopted and on its advantages: the time-to-event approach allows to predict the time to dropout, while the frailty and the time-varying covariates allow to fit the data and their complexity. The first aspect is relevant since it represents clear insights for universities and program managers, who can effectively use these predictions to intervene on time. In our case, dropout mainly occurs at the end of every year, but particularly after the first one. This means that students face difficulties especially at the beginning of their career. Potential reasons could be found in the low pre-academic preparation or in a misalignment in students’ expectations about university career. In this perspective, empowering the selection procedure and enriching the set of student information collected at the time of enrolment would help in providing more accurate and timely predictions. The second key takeaways relates to the characteristics of the most resilient (and, on the contrary, the most at risk) students. Girls, study grant recipients, and offsite students are those who retain more than their counterparts. The interpretation could be found in their (expected) higher motivation. Females are less represented in STEM disciplines, students with study grants probably feel the responsibility (and duty) of having received this opportunity, and offsite students have moved to another city – the one with the highest rents in Italy – probably thanks to the sacrifices of their families. Especially for female students, their resilience shows up late in the academic career (after 3 semester, half-way for graduation). As also confirmed in literature (Tinto, 2017), the motivation represents the main latent factor related to students' retention.

The last consideration relates to the adoption of an Early Warning System for detection of students at risk of dropout. This paper aims at setting the stage for a discussion about the timing of predictions as the result of an optimization problem to balance their accuracy and their timeliness. Findings indicate that as the student's career progresses, predictions' precision improves (as expected), but waiting for too long may lead the university to not have enough time to retain students. Evidence suggests that a possible optimal moment for prediction is the end of the first year, since the improvement in accuracy for the following semesters is nearly negligible.

Possible and interesting further development directions regard two main aspects. The former concerns the investigation of the heterogeneity at the degree programs level. Indeed, the dropout dynamics across degree courses might differ across time (e.g., the baseline hazard function of a degree program might be higher during the first year but lower during the second, with respect to the average), and time-invariant frailties are not able to catch this source of variability. Developing Cox models with time-varying frailties and degree program-specific parameters of covariates would significantly help the research in this direction. The latter regards the possibility to enrich the student-level dataset by including information about student motivation, psychological and personal aspects that would help the prediction allowing for even earlier accurate estimates.

Acknowledgements

Authors express special thanks to Mirko Giovio, who started the research and brilliantly graduated discussing it.

References

Alban, M., & Mauricio, D. (2019). Predicting university dropout through data mining: A systematic literature. *Indian Journal of Science and Technology*, *12*(4), 1–12.

ANVUR. (2018). *Rapporto biennale sullo stato del sistema universitario e della ricerca.* Retrieved from https://www.anvur.it/wp-content/uploads/2018/ 11/ANVUR-Completo-con-Link.pdf

Arulampalam, W., Naylor, R. A., & Smith, J. P. (2004). A hazard model of the probability of medical school drop-out in the uk. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *167*(1), 157–178.

Barragan, S., González, L., & Calderón, G. (2022). Modelling student dropout risk using survival analysis and analytic hierarchy process for an undergraduate accounting program. *Interchange*, 1–21.

Booth, L. L., & Satchell, S. E. (1995). The hazards of doing a phd: an analysis of completion and withdrawal rates of british phd students in the 1980s. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *158*(2), 297–318.

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., & Paganoni, A. M. (2022). Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques. *Studies in Higher Education*, *47*(9), 1935-1956.

Castro-Lopez, A., Cervero, A., Galve-González, C., Puente, J., & Bernardo, A. B. (2022). Evaluating critical success factors in the permanence in higher education using multi-criteria decision-making. *Higher Education Research & Development*, *41*(3), 628–646.

Chen, C., Sonnert, G., Sadler, P. M., Sasselov, D., & Fredericks, C. (2020). The impact of student misconceptions on student persistence in a MOOC. *Journal of Research in Science Teaching*, *57*(6), 879–910.

David, G. K., & Mitchel, K. (2012). *Survival analysis: a Self-Learning text*. Spinger.

De Valero, Y. F. (2001). Departmental factors affecting time-to-degree and completion rates of doctoral students at one land-grant research institution. *The Journal of higher education, 72*(3), 341–367.

Grove, W. A., Dutkowsky, D. H., & Grodner, A. (2007). Survive then thrive: determinants of success in the economics ph. d. program. *Economic Inquiry, 45*(4), 864–871.

Gury, N. (2011). Dropping out of higher education in France: a micro-economic approach using survival analysis. *Education Economics, 19*(1), 51–64.

Hegde, V., & Prageeth, P. P. (2018). Higher education student dropout prediction and analysis through educational data mining. In *2018 2nd international conference on inventive systems and control (ICISC)* (p. 694-699). doi: 10.1109/ICISC.2018.8398887

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association, 53*(282), 457–481.

Kehm, B. M., Larsen, M. R., & Sommersel, H. B. (2019). Student dropout from universities in Europe: A review of empirical literature. *Hungarian Educational Research Journal*, *9*(2), 147–164.

Klein, J. (1992). Semiparametric estimation of random effects using the cox model based on the EM algorithm. *Biometrics, 48*(3), 795–806. Retrieved 2022-04-06, from http://www.jstor.org/stable/2532345

Kleinbaum, D. G., & Klein, M. (1996). *Survival analysis a self-learning text.* Springer.

Lesik, S. A. (2007). Do developmental mathematics programs have a causal impact on student retention? an application of discrete-time survival and regression- discontinuity analysis. *Research in Higher Education, 48*(5), 583–608.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep, 50*, 163–170.

Min, Y., Zhang, G., Long, R. A., Anderson, T. J., & Ohland, M. W. (2011). Nonparametric survival analysis of the loss rate of undergraduate engineering students. *Journal of*

Engineering Education, *100*(2), 349–373.

Mussida, P. and Lanzi, P. L. (2020). A computational tool for engineer dropout prediction. *IEEE Global Engineering Education Conference*, pp. 1571-1576.

No, F., Taniguchi, K., & Hirakawa, Y. (2016). School dropout at the basic education level in rural Cambodia: Identifying its causes through longitudinal survival analysis. *International Journal of Educational Development*, *49*, 215–224.

Patacsil, F. F. (2020). Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models. *Universal Journal of Educational Research*, *8*(9), 4036–4047.

Plank, S. B., DeLuca, S., & Estacion, A. (2008). High school dropout and the role of career and technical education: A survival analysis of surviving high school. *Sociology of Education*, *81*(4), 345–370.

Rondeau, V., Gonzalez, J., Mazroui, Y., Mauguen, A., Diakite, A., Laurent, A., … Sofeu, C. (2019). Frailty pack: General frailty models: Shared, joint and nested frailty models with prediction; evaluation of failure-time surrogate endpoints. rpackage version 3.0.3. Retrieved from https://cran.r-project.org/package=frailtypack

Seidel, E., & Kutieleh, S. (2017). Using predictive analytics to target and improve fi year student attrition. *Australian Journal of Education*, *61*(2), 200–218.

Soares, T. M., Fernandes, N. d. S., Nóbrega, M. C., & Nicolella, A. C. (2015). Fac- tors associated with dropout rates in public secondary education in minas Gerais. *Educação e Pesquisa*, *41*, 757–772.

Spitzer, M. W. H., Gutsfeld, R., Wirzberger, M., & Moeller, K. (2021). Evaluating students' engagement with an online learning environment during and after covid- 19 related school closures: A survival analysis approach. *Trends in Neuroscience and Education*, *25*, 100168.

Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., & Raykar, V. C. (2007). On ranking in survival analysis: Bounds on the concordance index. In J. Platt, D.

Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems* (Vol. 20). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2007/file/

Thaithanan, J., Thinnukool, O., Chaichana, M., & Wanishsakpong, W. (2021). Using survival analysis to investigate undergraduate student dropout rates in the College of Arts, Media and Technology, Chiang Mai University. *Multicultural Education*, *7*(10).

Therneau, T. M., Grambsch, P. M., Therneau, T. M., & Grambsch, P. M. (2000). *The Cox Model* (pp. 39-77). Springer New York.

Tinto, V. (2017). Through the eyes of students. *Journal of College Student Retention: Research, Theory & Practice*, *19*(3), 254–269.

Utami, S., Winarni, I., Handayani, S. K., & Zuhairi, F. R. (2020). When and who dropouts from distance education? *Turkish Online Journal of Distance Education*, *21*(2), 141–152.

Vallejos, C. A., & Steel, M. F. (2017). Bayesian survival modelling of university outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(2), 613–631.

Van Der Haert, M., Arias Ortiz, E., Emplit, P., Halloin, V., & Dehon, C. (2014). Are dropout and degree completion in doctoral study significantly dependent on type of financial support and field of research?. *Studies in Higher Education*, *39*(10), 1885-1909.

Weybright, E. H., Caldwell, L. L., Xie, H., Wegner, L., & Smith, E. A. (2017). Predicting secondary school dropout among South African adolescents: A survival analysis approach. *South African journal of education*, *37*(2), 1–11.

Xie, Z. (2019). Modelling the dropout patterns of MOOC learners. *Tsinghua Science and Technology*, *25*(3), 313–324.

Appendices

Appendix A1. Students' distribution within programs

Table A1a: Distribution of students within the 16 programs and relative dropout percentage.

Degree Program Code	Number of students	% dropout
DG1	4,392	11.04
DG2	1,208	11.04
DG3	2,265	11.03
DG4	4,374	12.05
DG5	2,112	12.64
DG6	1,767	13.02
DG7	1,207	7.71
DG8	1,586	15.38
DG9	1,002	16.67
DG10	3,948	13.52
DG11	1,635	10.95
DG12	6,659	11.61
DG13	6,719	16.55
DG14	6,020	12.09
DG15	2,486	11.06
DG16	2,121	12.21

Appendix A2. Dropout risk across time by students' characteristics

From KM curves in Figure A2a, we observe that, despite the number of males is widely larger than the number of females (77.5% vs 22.5%), males are more likely to drop out. Regarding the family income, students with administrative support (*SG* category) are less likely to drop out across time. This could depend from the fact that students with SG are more motivated and feel the responsibility for having obtained a grant. The highest risk category, especially right after the end of first semester, is that of high income students. Nonetheless, on the long term, also low income students show a higher dropout probability with respect to the other categories, which suggests that students with a more disadvantaged background, who do not receive administrative support, are more exposed to dropout, especially on the long term. The dropout probability of the Medium category reaches results very close to the SG group at the end of the follow up time. For what concerns student origins, Offsite students (i.e., students coming from other regions who moved to Milan to study at PoliMi) have on average a lower dropout risk with respect to Milanese and Commuter students. Regarding the type of high school attended before the enrolment at PoliMi, most of the students come from Scientific school (80.5%) and result to be the ones with lowest dropout risk. Students coming from Classical schools present a significant higher risk of dropout during the first year, while, on the opposite, students who attended a high school abroad are less likely to dropout at the beginning but more likely to dropout during their third year. At the end of the follow-up, technical schools and all other types of high schools show a relatively high dropout probability. Furthermore, also the high school grade results to be a determinant of the dropout risk. We identify 75 as the threshold that differentiates the most the two populations, highlighting that students with a high school final mark lower than 75 have on average higher risk of dropout with respect to the others. Students enrolling at PoliMi later than the standard age (19), tent to drop more than younger students, especially after the first year.

Focusing on the early performance at PoliMi (KM curves in Figure A 2b), we observe a lower dropout risk for those students obtaining an admission score higher than 71, that resulted to be the most significant threshold. In terms of ECTS and GPA measured at the end of first semester, we observe that obtaining less than 10 ECTS is an extremely predictive risk factor. The sharp difference between the two KM curves highlights the importance of this information and its predictive power. Among the students who obtained at least 10 ECTS, having a GPA lower than 22, i.e., the most discriminant value, constitutes a further risk factor. Lastly, given our interest in investigating the difference across degree programs, we observe that the 16 KM curves show heterogeneous dropout dynamics across faculties, detecting up to a 13% difference in the dropout percentage at the end of follow-up across faculties.

Figure A2a: Kaplan-Meier Curves for Gender Income, Origins, HighschoolType, High schoolGrade, and AdmissionAge.

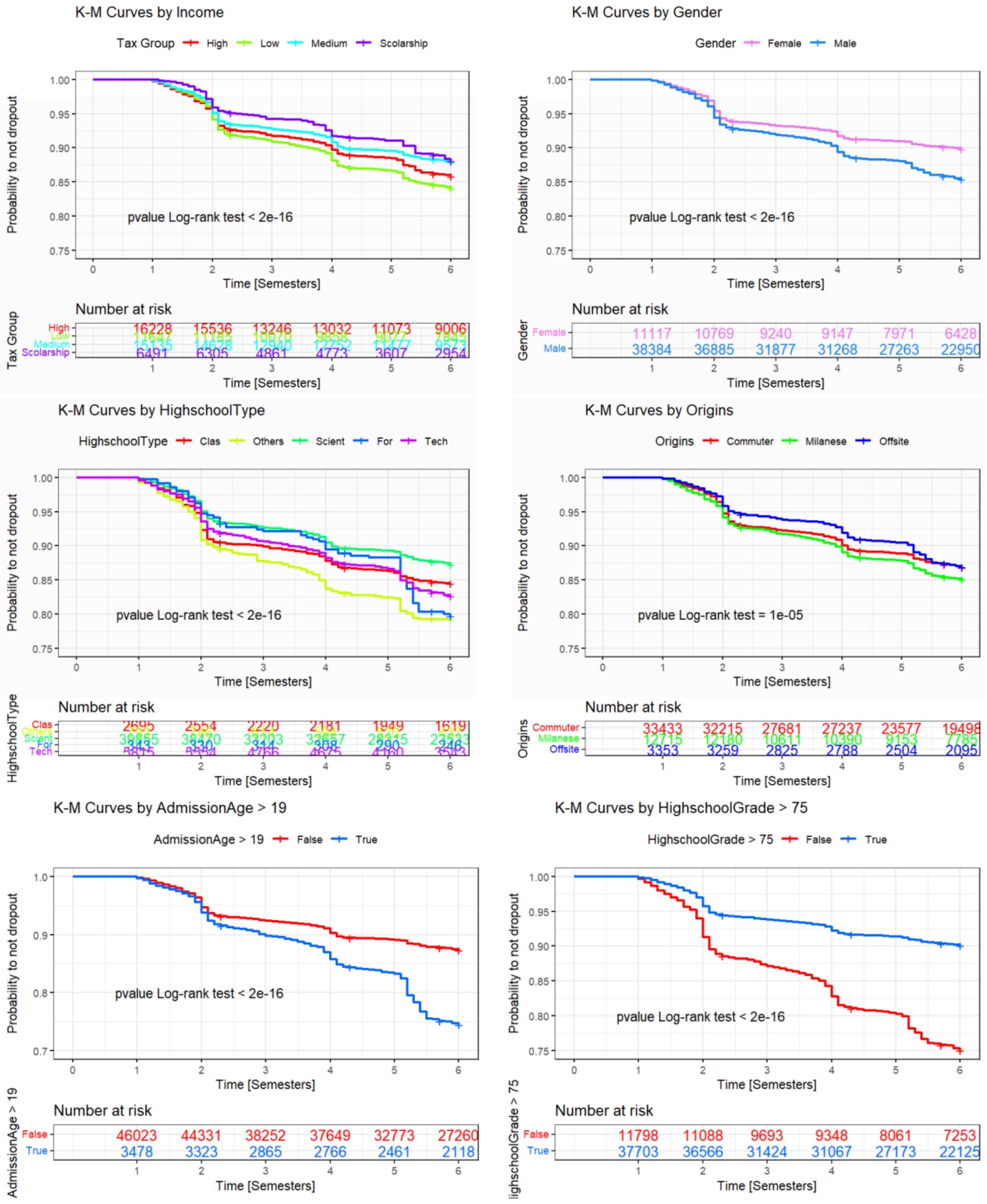
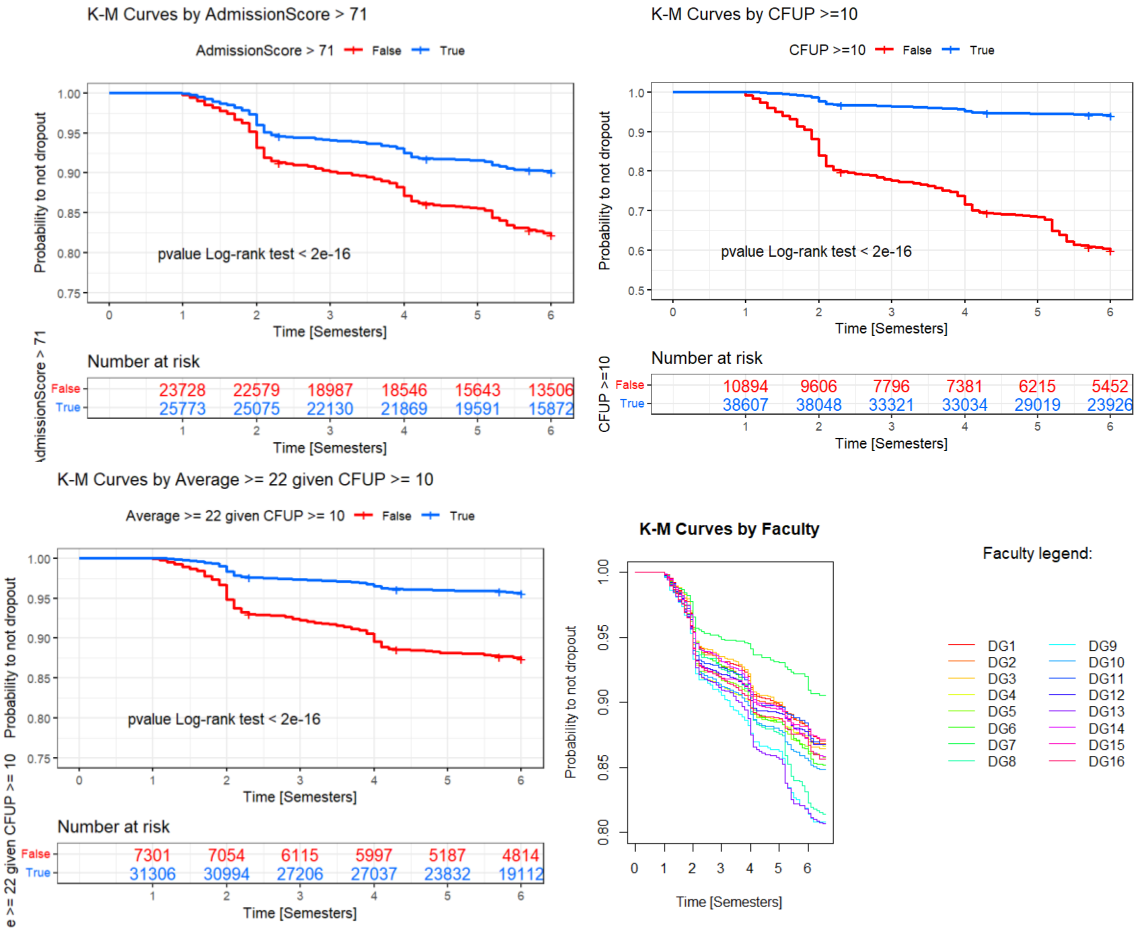


Figure A2b: Kaplan-Meier Curves for *AdmissionScore*, *ECTS*, *GPA*, and *Degree Program*.



Note: For each numerical covariate, the threshold represents the value for which the difference between the two Kaplan-Meyer curves is maximized.

Appendix A3. Baseline survival and hazard functions of the shared frailty Cox models

Figure A3a: Estimated baseline survival and hazard functions of the shared frailty Cox model with time-invariant covariates.

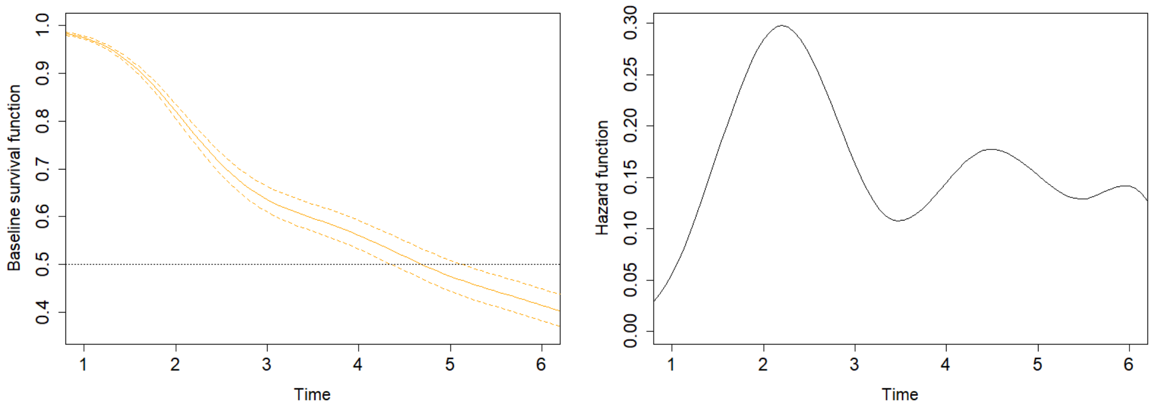
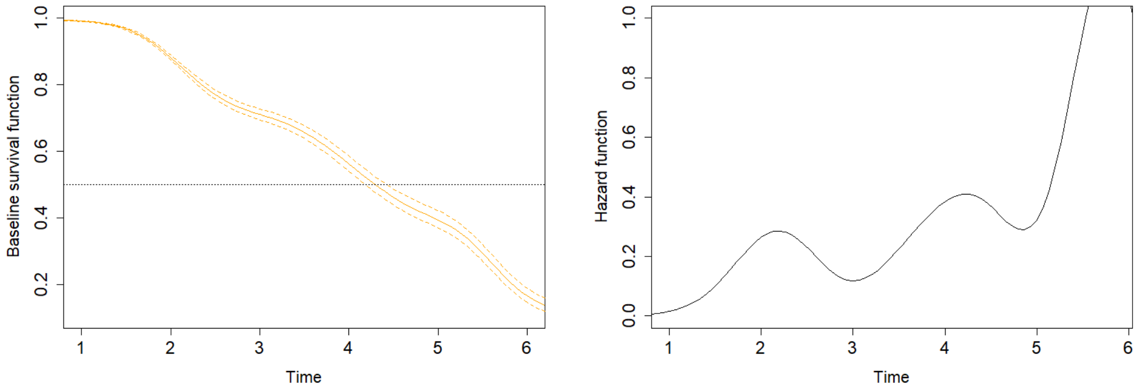


Figure A3b: Estimated baseline survival and hazard functions of the shared frailty Cox model with time-varying covariates.



MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

12/2023 Masci, C.; Cannistrà, M.; Mussida, P.
Modelling time-to-dropout via Shared Frailty Cox Models. A trade-off between accurate and early predictions

11/2023 Gatti, F.; Perotto, S.; de Falco, C.; Formaggia, L.
A positivity-preserving well-balanced numerical scheme for the simulation of fast landslides with efficient time stepping

10/2023 Corti, M.; Antonietti, P.F.; Bonizzoni, F.; Dede', L., Quarteroni, A.
Discontinuous Galerkin Methods for Fisher-Kolmogorov Equation with Application to Alpha-Synuclein Spreading in Parkinson's Disease

09/2023 Buchwald, S.; Ciaramella, G.; Salomon, J.
Gauss-Newton oriented greedy algorithms for the reconstruction of operators in nonlinear dynamics

08/2023 Bonizzoni, F.; Hu, K.; Kanschat, G.; Sap, D.
Discrete tensor product BGG sequences: splines and finite elements

06/2023 Artoni, A.; Antonietti, P. F.; Mazzieri, I.; Parolini, N.; Rocchi, D.
A segregated finite volume - spectral element method for aeroacoustic problems

07/2023 Garcia-Contreras, G.; Còrcoles, J.; Ruiz-Cruz, J.A.; Oldoni, M; Gentili, G.G.; Micheletti, S.; Perotto, S.
Advanced Modeling of Rectangular Waveguide Devices with Smooth Profiles by Hierarchical Model Reduction

05/2023 Fumagalli, I.; Vergara, C.
Novel approaches for the numerical solution of fluid-structure interaction in the aorta

04/2023 Quarteroni, A.; Dede', L.; Regazzoni, F.; Vergara, C.
A mathematical model of the human heart suitable to address clinical problems

03/2023 Africa, P.C.; Perotto, S.; de Falco, C.
Scalable Recovery-based Adaptation on Quadtree Meshes for Advection-Diffusion-Reaction Problems