Determinants of Bachelor's Degree Completion: The Role of Pre-University Background and Academic Progress

Francesca Pierri[1(✉)] and Chrys Caroni[2]

[1] Department of Economics, University of Perugia, Perugia, Italy
francesca.pierri@unipg.it
[2] Department of Mathematics, National Technical University of Athens, Athens, Greece
ccar@math.ntua.gr

Abstract. Universities can improve their training systems by systematically analyzing students' academic trajectories to better understand progression patterns and identify early risk factors for dropout. Administrative data—such as exams taken, credits earned, grades, and time to completion—provide valuable information for this purpose.

In this study, a cohort of students enrolled in three-year degree programs at the University of Perugia was followed over a ten-year period (2015–2025), exceeding the nominal duration of the degree in order to capture final academic outcomes. Multinomial logistic regression and competing risks models based on the Fine–Gray approach were applied to identify the main determinants of academic success and dropout.

The results show that, while pre-university background characteristics are relevant at entry, their predictive power decreases over time. Academic performance indicators—particularly accumulated and remaining credits—emerge as the main drivers of both graduation and attrition. The comparison between models highlights a trade-off between predictive accuracy and early interpretability, with time-to-event approaches providing a more appropriate representation of student trajectories.

These findings support the importance of early monitoring of academic performance to identify at-risk students and design targeted support strategies, contributing to more effective and data-driven university policies.

Keywords: university dropout · logistic regression · survival analysis · competing risks

1 Introduction

The analysis of students' academic trajectories has become a central topic in higher education research, as universities increasingly rely on empirical evidence to improve teaching effectiveness, support student success, and reduce

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 3–19, 2027.
https://doi.org/10.1007/978-3-032-30521-3_1

dropout rates. Monitoring students' progress through administrative data—such as exams taken, credits accumulated, grades obtained, and time to completion—provides institutions with valuable insights into how students navigate their academic pathways. These data allow universities to identify patterns in performance, detect potential difficulties early in the academic career, and design targeted interventions aimed at improving persistence and degree completion.

Student persistence and graduation outcomes have long been studied within the broader literature on higher education. Foundational theoretical frameworks emphasize the role of both individual characteristics and institutional environments in shaping students' educational trajectories. In particular, Tinto's model of student integration highlights how academic and social integration within the university environment influence persistence decisions and the likelihood of completing a degree [19,20]. Similarly, Bean [1] proposed a model inspired by organizational turnover theory, emphasizing how students' attitudes, academic performance, and institutional experiences affect their decisions to remain enrolled. These theoretical contributions have provided the basis for extensive empirical research investigating the determinants of student success and dropout.

More recent studies confirm that university dropout remains a significant challenge for higher education systems worldwide. Recent reviews estimate that nearly one in five students does not complete a university degree, with substantial consequences for individuals, institutions, and society. Non-completion is associated with poorer labour market outcomes and increased economic vulnerability, highlighting the importance of identifying the factors that influence persistence and degree completion [14].

A substantial strand of literature focuses on the role of pre-university background in shaping academic outcomes. Socio-economic status, parental education, prior academic performance, and the type of secondary school attended are consistently identified as key determinants of student success [3,6,14]. Students from disadvantaged socio-economic backgrounds often face structural barriers that increase the probability of dropout or delay in completing their studies. Empirical evidence shows that lower family cultural and economic capital significantly increases the risk of dropout, suggesting that inequalities accumulated before university continue to influence outcomes during higher education [15].

In addition to background characteristics, research increasingly highlights the importance of academic progress during university as a key predictor of final outcomes. Indicators such as credit accumulation, failed examinations, grade point average, and course participation provide strong signals of students' academic trajectories. Beyond initial background characteristics, an increasing number of studies highlight the importance of academic progress during the early stages of university as a key predictor of long-term success. Early academic performance, credit accumulation, and timely completion of examinations are often strong indicators of eventual graduation [8,12]. Students who fall behind in the first years of their studies frequently face increasing difficulties in catching up, which can ultimately lead to prolonged enrollment or dropout. Moreover, recent predictive studies show that accumulated credits and failed courses are among the

most powerful predictors of university dropout and academic success, with their importance evolving throughout the academic career [21]. Early academic performance is therefore often considered a critical indicator for identifying students at risk of falling behind or abandoning their studies.

The growing availability of administrative and digital learning data has led to the emergence of new research fields such as learning analytics and educational data mining, which aim to transform educational data into actionable knowledge. These approaches apply statistical and machine learning techniques to analyze student behavior, predict academic outcomes, and support evidence-based decision-making in higher education institutions. Recent literature highlights the increasing use of predictive analytics and artificial intelligence to identify students at risk of dropout and to design early intervention strategies aimed at improving retention and academic success [9].

Within this context, longitudinal and event-history approaches have become particularly relevant for studying students' academic careers. Survival analysis and competing risks models allow researchers to analyze the timing of events such as graduation, dropout, or prolonged enrollment while accounting for the dynamic nature of academic pathways. These methods provide a more nuanced understanding of student progression and allow institutions to identify critical periods during which the risk of dropout is highest [7].

Building on this growing literature, the present study examines the determinants of bachelor's degree completion by analyzing the academic trajectories of students enrolled in three-year programs at the Faculty of Economics of the University of Perugia. Using administrative data tracking students from enrollment over a ten-year period (2015–2025), the analysis observes academic careers beyond the nominal duration of the degree in order to capture different possible outcomes, including timely graduation, delayed completion, or non-completion.

The empirical analysis applies two complementary methodological approaches: multinomial logistic regression and competing risks survival models. The first method allows us to classify students according to different academic outcomes and identify the factors associated with each trajectory. The second approach models the timing of events during the academic career while accounting for the presence of multiple possible exit states. By comparing these two methodologies, the study highlights their respective advantages in terms of interpretability and their usefulness for transforming administrative data into predictive knowledge.

Overall, the objective of the paper is twofold. First, we aim to identify the main determinants of bachelor's degree completion by considering both students' pre-university background and indicators of academic progress during university studies. Second, we compare two statistical approaches widely used in the analysis of educational data to evaluate their effectiveness in modeling academic trajectories and supporting data-driven policies in higher education. Through this analysis, the study contributes to the understanding of students' academic pathways and provides empirical evidence that may support universities in designing policies aimed at promoting academic success and timely graduation.

The remainder of this paper is organized as follows. Section 2 outlines the methodological framework and describes the statistical techniques employed in the analysis. Section 3 presents the empirical application, including a description of the dataset and the main results. Section 4 concludes with a discussion of the findings and their implications.

2 Methodology

Statistical and machine learning methods can help create predictive models that estimate the chances of graduating on time, graduating late, or dropping out. Some of the most popular methods include Logistic Regression, Decision Tree Learning, Random Forests, and Survival Analysis, which is used to predict the time until graduation or dropout.

Here multinomial logistic regression and competing risks models based on the Fine–Gray approach were applied to consider the different cause of dropping out. Descriptive statistics were used before developing the models. Model selection was performed on a training dataset comprising 80% of the observations, randomly drawn without replacement from the full sample. The fitted model was then evaluated on the remaining 20% of the data, serving as an out-of-sample validation set, to assess predictive performance and generalizability. This approach allowed us to quantify the model's accuracy and robustness, ensuring that the estimated relationships were not driven by overfitting to the training data.

All analyses were primarily conducted using SAS® software [17,18], which was employed for the estimation of multinomial logistic regression models and Fine–Gray competing risks models through the `PROC LOGISTIC` [17] and `PROC PHREG` [18] procedures. To complement the inferential analysis and to evaluate the predictive performance of the competing risks model, additional analyses were performed in the R statistical environment. In particular, the `timeROC` package [4,5] was employed to compute time-dependent receiver operating characteristic curves and corresponding area under the curve (AUC) measures.
This complementary use of R allowed for a more detailed assessment of model discrimination over time, which is not directly provided within the standard SAS implementation for competing risks models. In this way, SAS ensured the estimation of the regression framework, while R was used for advanced predictive performance evaluation and graphical diagnostics.

2.1 Multinomial Logistic Regression

Multinomial logistic regression is basically an extension of binary logistic regression that looks at how a set of explanatory variables relates to a categorical dependent variable with more than two unordered outcomes. It is part of the family of discrete choice models and works well when the response variable has several mutually exclusive options that do not follow any natural order. Thanks to its flexibility in dealing with nominal outcomes and multiple options, multinomial logistic regression is commonly used in areas like economics, sociology, political science, marketing, and education, especially for research on choice behavior,

career paths, or educational trajectories. Multinomial logistic regression with a nominal dependent variable can be modeled using the generalized logit (glogit) link, which represents the log-odds of each possible outcome compared to a chosen reference category.

Let Y be a categorical variable with J unordered categories ($j = 1, 2, ..., J$) and let $X = (X_1, X_2, ..., X_p)$ denote a vector of explanatory variables. By convention, one category (e.g., category J) is chosen as the baseline. The generalized logit model then specifies the log-odds of being in category j relative to the reference category J as a linear function of the predictors:

$$\ln \frac{P(Y = j \mid X)}{P(Y = J \mid X)} = \beta_{j0} + \beta_{j1}X_1 + \cdots + \beta_{jp}X_p, \quad j = 1, \ldots, J-1 \tag{1}$$

From this, the predicted probabilities for each category are:

$$P(Y = j \mid X) = \frac{\exp(\beta_{j0} + \beta_{j1}X_1 + \cdots + \beta_{jp}X_p)}{1 + \sum_{k=1}^{J-1} \exp(\beta_{k0} + \beta_{k1}X_1 + \cdots + \beta_{kp}X_p)}, \quad j = 1, \ldots, J-1 \tag{2}$$

$$P(Y = J \mid X) = \frac{1}{1 + \sum_{k=1}^{J-1} \exp(\beta_{k0} + \beta_{k1}X_1 + \cdots + \beta_{kp}X_p)} \tag{3}$$

Here, each coefficient β_{jk} represents the effect of predictor X_k on the log-odds of being in category j relative to the reference category J. The model ensures that predicted probabilities are positive and sum to 1.

Parameters were estimated via maximum likelihood. Multicollinearity among covariates was assessed prior to model estimation. Variable selection was conducted using a stepwise procedure based on Wald statistics as an exploratory tool to identify a parsimonious set of predictors. Model selection was further supported by likelihood-based criteria, including the Akaike Information Criterion (AIC), the Schwarz Bayesian Criterion (SBC), and the -2 Log Likelihood, as provided by SAS output. Given the well-known limitations of stepwise procedures, including potential model instability and inflated Type I error rates, results were interpreted with caution and emphasis was placed on consistency across model specifications. Model fit was assessed using standard goodness-of-fit measures, including the Hosmer-Lemeshow test [13], while predictive performance was evaluated through the Brier Score.

2.2 Competing Risks Analysis

Survival models are often used in risk analysis to estimate how long it might take for one or more events to happen, for example death, failure of a machine, or relapse of a disease. This kind of study is also called time-to-event analysis or event history analysis. The survival function gives the probability that the event occurs after time t,

$$S(t) = P(T > t) \tag{4}$$

where T is the random variable representing the time to the event. The survival function is related to the cumulative distribution function $F(t)$, that represents the total chance that an event will happen within time t.

$$S(t) = 1 - F(t). \tag{5}$$

A particular interest lies in the hazard function, that is the instantaneous risk that the event occurs at time t, given survival until time t

$$h(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t | T > t]}{\delta t} = \frac{f(t)}{S(t)}\,. \tag{6}$$

Competing risks analysis deals with situations in which an observational unit is at risk of experiencing one of several mutually exclusive events, called competing events and the occurrence of one event precludes the occurrence of the others. Two main approaches exist in this context: the cause-specific hazard model and the subdistribution hazard model (Fine–Gray). The first analyzes the hazard of a particular event by treating other competing events as censored observations. We define the cause-specific hazard for the kth of K causes as

$$h_k(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t,\ D = k | T > t]}{\delta t}\,, \tag{7}$$

where D denotes the cause of failure. It models the instantaneous risk that the kth event occurs at time t, given that no event has occurred yet. However, it does not directly model the cumulative incidence of the event.

In this context we used the Fine–Gray approach [10,11,16] because it directly models the subdistribution hazard, which is linked to the cumulative incidence function (CIF) of a specific event in the presence of competing risks. In this case, individuals who have experienced one of the competing events remain in the risk set with special weighting rather than being censored. The subdistribution hazard function [10] estimates the risk of failure from cause k,

$$h_k^s(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t, D = k \mid T \geq t \text{ or } (T \leq t \text{ and } D \neq k)]}{\delta t} \tag{8}$$

and expresses the instantaneous rate of failure at time t among subjects that have not yet experienced an event of type k.

Model specification was guided by likelihood-based criteria (AIC) and by the magnitude and precision of subdistribution hazard ratios and their confidence intervals, with the aim of obtaining a parsimonious and interpretable model. The proportional hazards assumption was assessed for all covariates using standard diagnostic procedures. Evidence of non-proportionality was observed for the variable graduation mark. To address this issue, an interaction term between graduation mark and the logarithm of time was introduced and included in the model alongside the original graduation mark variable. This time-dependent specification allows the effect of graduation mark on the hazard to vary over time, providing a flexible adjustment for the non-proportionality while preserving interpretability of the covariate's effect on the outcome.

To assess model performance, both calibration and discrimination were evaluated. Calibration was examined through the comparison of cumulative incidence functions (CIFs) between training and validation samples. Discrimination was assessed by computing time-dependent area under the curve (AUC) measures within a competing risks framework. This approach allowed the evaluation of the model's ability to distinguish between different event types over time. Overall, this combined strategy provides a comprehensive assessment of model performance, integrating regression-based inference with predictive validation.

3 Application

3.1 Data

A cohort of students enrolled in 2015 in the three bachelor degree programs in Economics at the University of Perugia was followed for ten years. Each student was identified by an anonymous identification code. Along with the details collected during registration, such as gender, residence, secondary school attended, diploma, and graduation mark, information about their university career was also provided. Therefore, for each student, we have details like the exams they have taken, the training credits earned, the exam dates, the grades received, and their current status: graduated (status = 1), voluntary withdrawal from studies (status = 2), withdrawal due to inactivity (status = 3), or still active (status = 0).

Table 1. Distribution of students' current status by course

Degree Code		Status				
		0	1	2	3	Total
L021	N	24	209	79	64	376
L025	N	15	40	13	9	77
L026	N	3	19	6	13	41
Total	N	42	268	98	86	494
	%	8.5	54.25	19.84	17.41	100

The three bachelor degrees in Economics are located in different Umbria Region districts and are identified with codes: L021 (Business Administration in Perugia), L025 (Business Administration in Terni) and L026 (Business for Tourism in Assisi). The distribution of students by status and degree course (Table 1) shows that the highest percentage of students attends the L021 course (76.11%). Moreover, keeping track of how long it takes students to finish their studies, we notice (Table 2) that just 54.25% completed their degree program. Among graduates, 37.31% finished within the expected duration of three years, 41.79% completed in four years, 14.18% in five years, while 6.72% required between

Table 2. Distribution of students by status and time

		Time										
Status		1	2	3	4	5	6	7	8	9	10	Total
0	N	0	0	0	0	0	0	0	0	0	42	42
	%	0	0	0	0	0	0	0	0	0	8.5	8.5
1	N	0	0	100	112	38	8	6	3	1	0	268
	%	0	0	20.24	22.67	7.69	1.62	1.21	0.61	0.2	0	54.25
2	N	54	18	7	4	7	4	2	1	1	0	98
	%	10.93	3.64	1.42	0.81	1.42	0.81	0.4	0.2	0.2	0	19.84
3	N	0	0	2	2	7	3	63	5	4	0	86
	%	0	0	0.4	0.4	1.42	0.61	12.75	1.01	0.81	0	17.41
Total	N	54	18	109	118	52	15	71	9	6	42	494
	%	10.93	3.64	22.06	23.89	10.53	3.04	14.37	1.82	1.21	8.5	100

Table 3. Distribution of type of school attended by gender.

School		Male	Female	Total
Technical	N	120	86	206
	%	43.17	39.81	41.70
Scientific	N	103	43	146
	%	37.05	19.91	29.55
Classic	N	23	44	67
	%	8.27	20.37	13.56
Professional	N	20	18	38
	%	7.19	8.33	7.69
Mixed*	N	1	20	21
	%	0.36	9.26	4.25
Outside Italy	N	11	5	16
	%	3.96	2.31	3.24
Total	N	278	216	494

*Includes elements of classical and scientific

six and ten years. Students belonging to status 2 (19.84%), closed their university position mainly in the first (55.10%) and second (18.37%) years. A further 17.41% of the students failed to complete their studies (status 3); they were removed for inactivity by the system mainly at the end of the seventh year (73.26%).

Students signing up for Economics courses live mainly in the Umbria region (86.23%) and attended a Technical (41.7%), Scientific (29.55%) or Classical (13.56%) high school (Table 3).

The quantitative variables used in the analysis are shown along with their distributions in Table 4.

Table 4. Data distribution

Variable	N	Minimum	Maximum	Median	Mean	Std Dev
School Graduation Mark	492	0	100	76.00	76.30	16.37
Exam Grade Year 1 (mean 1)	404	18	30	24.00	23.90	3.10
Exam Grade Year 2	332	18	30	24.98	24.70	2.76
Exam Grade Year 3	318	18	30	25.89	25.48	2.59
Exam Grade Year 4	200	18	30	23.65	23.78	2.90
Exam Grade Year 5	72	18	30	23.00	23.23	3.30
Total exams completed	494	0	26	23.00	15.47	10.86
Total exams left	494	0	22	0.00	4.37	5.90
Exams completed year 1	494	0	10	4.00	4.18	2.76
Exams completed year 2	494	0	11	5.00	3.96	3.34
Exams completed year 3	494	0	15	5.00	4.68	4.11
Exams completed year 4	494	0	14	0.00	1.92	2.79
Exams completed year 5	494	0	11	0.00	0.51	1.44
Training Credits Completed year 1	494	0	78	36.00	33.49	22.14
Training Credits Completed year 2	494	0	81	34.50	30.10	25.18
Training Credits Completed year 3	494	0	99	36.00	31.71	27.36
Training Credits Completed year 4	494	0	99	0.00	13.30	19.36
Training Credits Completed year 5	494	0	81	0.00	3.50	10.13
Training Credits Left Year 1	494	0	54	18.00	20.51	22.14
Training Credits Left Year 2	494	0	63	28.50	32.90	25.18
Training Credits Left Year 3	494	0	63	27.00	31.29	27.36
Training Credits Left after 2 years	494	0	117	45.00	53.41	44.77
Training Credits Left after 3 years	494	0	180	69.00	84.70	69.56

3.2 Results

Multinomial Logistic Regression. The multinomial logistic regression model was initially estimated using all available data on students' academic careers during the first five years of study. This time frame was chosen to capture a sufficiently long trajectory, although voluntary dropout typically occurs during the second year.

Table 5 tells us that as the number of training credits left at the fifth year increases, the probability of being in status 0 (OR=1.118) and especially in status 2 (OR = 1.174) and 3 (OR = 1.172) also increases. Moreover, comparing the

Table 5. Effects of statistically significant covariates in the multinomial logistic regression model based on the first five years of the university career. Odds Ratios (OR) and 95% Confidence Intervals (CI) are reported ($p < 0.05$).

	status = 0		status = 2		status = 3	
	OR	95% CI	OR	95% CI	OR	95% CI
Degree Course*						
L025	0.925	0.127 – 6.753	0.073	0.006 – 0.858	0.066	0.005 – 0.801
L026	5.298	0.598 – 46.949	2.034	0.086 – 48.215	7.563	0.356 – 160.836
Credits Left Year 5	1.118	1.078 – 1.161	1.174	1.127 – 1.224	1.172	1.125 – 1.222

Reference categories: status=1, * L021
Hosmer and Lemeshow Goodness-of-Fit Test p=0.3797

courses of study, course L021 is associated with a higher relative odds of failure than course L025. In fact, the odds ratios for status 2 (OR = 0.073, CI = 0.006-0.858) and status 3 (OR = 0.066, CI = 0.005–0.801) have values significantly less than 1. However, from Table 1 it is clear that most students in status 2 tend to drop out of their studies by their second year. Therefore, to evaluate whether early academic indicators could improve predictive performance, a second specification was estimated using academic data from only the first two years of study.

This approach allows the model (Table 6) to capture early signs of attrition and assess whether information from the initial years of university can enhance the prediction of different academic outcomes.

Table 6. Effects of statistically significant covariates in multinomial logistic regression considering among the other variables the first two years of the University career: Odds Ratio (OR) and Wald Confidence Interval (CI) are reported ($p < 0.05$).

	status = 0		status = 2		status = 3	
	OR	95% CI	OR	95% CI	OR	95% CI
Degree Course*						
L025	1.072	0.324 – 3.541	0.052	0.008 – 0.328	0.053	0.008 – 0.344
L026	1.015	0.204 – 5.059	0.194	0.023 – 1.678	0.786	0.112 – 5.513
Credit Left after 2 years	1.064	1.042 – 1.087	1.082	1.051 – 1.114	1.093	1.062 – 1.125
Exam Grade Year 2	0.996	0.906 – 1.093	0.846	0.762 – 0.939	0.884	0.798 – 0.978

Reference categories: status=1, * L021
Hosmer and Lemeshow Goodness-of-Fit Test p=0.9832

The model estimated using the first two years (Table 6) confirms persistent differences between the two degree programs, L025 and L021, particularly in the early identification of dropout risk. Holding other covariates constant, an increase in the number of credits remaining at the end of the first two years is

associated with a higher probability of dropout. Conversely, a higher weighted average of grades in the second year is negatively associated with the likelihood of experiencing outcomes corresponding to status 2 or 3, suggesting a protective effect of strong academic performance during the early stages of the degree program. Nevertheless, the Hosmer and Lemeshow Goodness-of-Fit test states the superiority of the first model (Table 5), and the Brier Score on the validation data backs it up as well (0.26 against 0.34).

Competing Risks. The Fine–Gray competing risks model (Table 7) provides a dynamic representation of academic trajectories accounting for the presence of multiple mutually exclusive outcomes. Results indicate that higher values of the covariates are associated with an increased subdistribution hazard of graduation (status 1), in particular higher weighted average grades in the third year (HR = 1.055, CI = 1.021–1.091) and higher secondary school graduation marks (HR = 1.131, CI = 1.096–1.167).

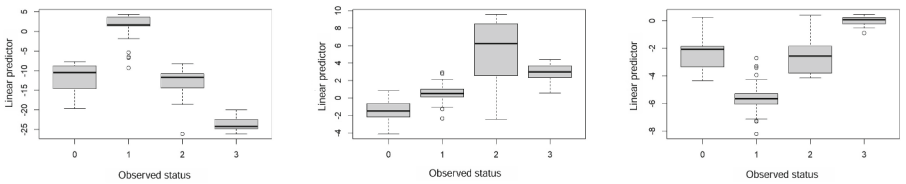
Table 7. Effect of covariates in Fine–Gray subdistributional hazard regression model: Subdistributional Hazard Ratio (HR) and 95% Confidence Interval (CI)

	status = 1		status = 2		status = 3	
	HR	95% CI	HR	95% CI	HR	95% CI
Degree Course*						
L025					0.609	0.259 – 1.436
L026					1.594	1.151 – 2.209
School Grad Mark	1.131	1.096 – 1.167	1.051	1.039 – 1.063	0.962	0.950 – 0.973
Time-by-Mark interact.	0.917	0.895 – 0.939	0.959	0.953 – 0.965	1.023	1.018 – 1.028
Mean Year 3	1.055	1.021 – 1.091				
Credits Compl. Year 1				0.974	0.956	0.991
Credits Compl. Year 3				0.959	0.934	0.986
Credits Compl. Year 4				0.936	0.891	0.982
Credits Left Year 2			1.082	1.065 – 1.101		
Credits Left Year 5	0.883	0.847 – 0.921				

Reference category: * L021

Conversely, it decreases as the number of remaining training credits in the fifth year increases (HR = 0.883, CI = 0.847–0.921), suggesting that delayed academic progress is associated with a lower likelihood of degree completion. For dropout (status 2), the subdistribution hazard increases with the number of credits sill to be completed in the second year (HR = 1.082, CI = 1.065–1.101) and with higher secondary school graduation marks (HR = 1.051, CI = 1.039–1.060). However, the effect of secondary school performance decreases over time, as indicated by a subdistribution hazard ratio below one for the corresponding interaction term with time. Regarding inactivity-related withdrawal (status 3), students enrolled in the L026° program exhibit a higher risk compared to those in L021 (HR =

1.594, CI =1.151–2.209). In contrast, a higher accumulation of credits in the first, third, and especially fourth years is associated with a reduced risk of this outcome (HR<1), indicating a protective effect of sustained academic progress. Secondary school graduation marks also show a protective association (HR = 0.962, CI= 0.950–0.973), although this effect varies over time. The boxplot analysis of the linear predictors provides additional insight into model discrimination (Fig. 1). For statuses 1 (graduation) and 3 (inactivity), a clear separation between outcome groups is observed, indicating strong discriminative ability. In contrast, status 2 (dropout) shows partial overlap between categories, reflecting the intermediate and heterogeneous nature of this outcome.



(a) Graduation (b) Dropout (c) Inactivity

Fig. 1. Distribution of the estimated event-specific linear predictors across observed outcome categories.

Table 8. Time-dependent AUC values of the Fine–Gray competing risks model on the validation set.

Event	Time (years)							
	2	3	4	5	6	7	8	9
Graduation (Status 1)	–	–	1.000	1.000	1.000	1.000	1.000	0.997
Dropout (Status 2)	1.00	1.00	1	0.988	0.984	0.987	0.966	0.965
Inactivity (Status 3)	–	–	–	0.871	0.897	0.976	0.977	0.980

The predictive performance of the Fine–Gray competing risks model was assessed on the validation sample using time-dependent AUC and the distribution of linear predictors across observed outcomes. The results (Table 8) show a generally high discriminative ability across all event types, although with different patterns over time.

For graduation (status 1), AUC values are close to 1 from year 4 onwards, indicating an almost perfect ability of the model to distinguish between students who graduate and those who do not at later stages of the academic trajectory. This result is consistent with the structure of the data, as no graduation events are observed in the early years.

For dropout (status 2), the model exhibits very high discrimination at early time points (AUC = 1 at years 2 and 3), followed by a slight decrease over time, with values stabilizing around 0.96–0.99. This pattern suggests that early dropout is highly predictable, while later dropout becomes more heterogeneous and therefore harder to distinguish.
For inactivity-related withdrawal (status 3), AUC values are not available in the initial years due to the absence of events, but show a progressive increase from year 5 onwards, reaching values above 0.97 in later periods. This indicates an improvement in predictive accuracy over time, likely reflecting the cumulative effect of academic performance indicators. These results should be interpreted in light of the relatively small number of events at specific time points, particularly in the early follow-up period. Overall, they confirm the strong discriminative ability of the model, while also highlighting differences in predictability across event types and time horizons.

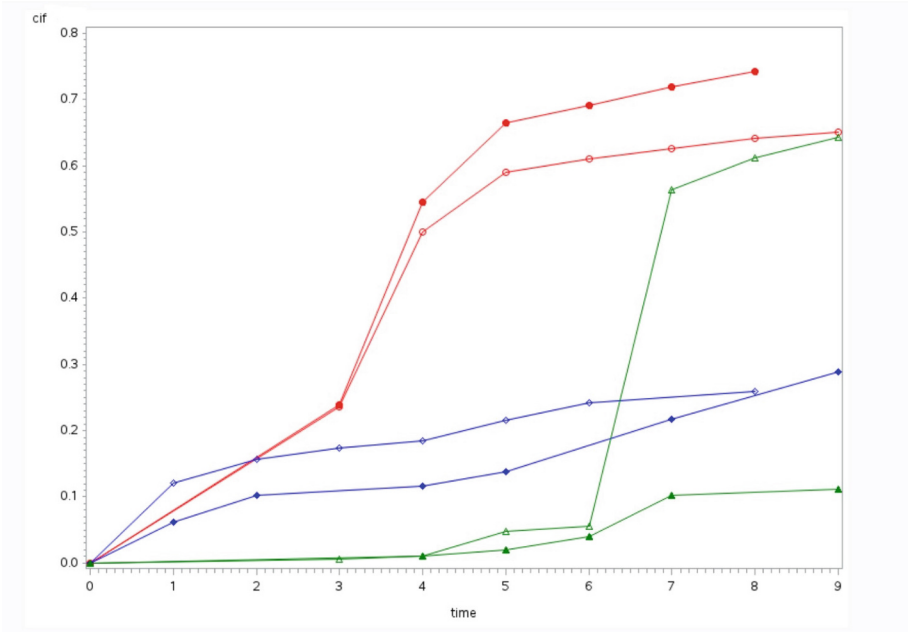


Fig. 2. Cumulative incidence functions (CIF) estimated for training (open symbols) and validation (filled symbols) samples under the Fine–Gray competing risks framework. Upper, intermediate, and lower curves correspond to Events 1 (red), 2 (blue), and 3 (green), respectively. (Color figure online)

Finally, the comparison of cumulative incidence functions between training and validation samples (Fig. 2) provides a qualitative assessment of model calibration over time. Good agreement is observed for graduation (status 1) and inactivity-related withdrawal (status 3) in the early follow-up period, while discrepancies

emerge at later times. For dropout, a systematic difference between training and validation curves suggests lower predictive stability for this outcome.

Overall, these results confirm that academic performance indicators are the strongest determinants of student trajectories, while pre-university characteristics mainly act as initial conditions whose effect diminishes over time. The Fine–Gray model further highlights the importance of accounting for competing risks and temporal dynamics in the analysis of academic progression.

4 Conclusions

This study extends previous research by integrating pre-university characteristics with longitudinal information on students' academic trajectories, enabling a direct comparison between baseline attributes and time-varying performance indicators. The results indicate that pre-university characteristics (e.g., secondary school graduation marks and school type) are informative at entry but their relative contribution to prediction decreases as students progress through their academic careers. This pattern is consistent with prior literature [2,20] and supports the interpretation of these variables as initial conditions, whereas university performance reflects a dynamic process of academic integration [19,20].

Empirically, academic performance indicators—particularly accumulated and remaining credits—emerge as the strongest predictors of both graduation and dropout. Once in-course performance is accounted for, the explanatory power of pre-entry characteristics is substantially reduced. Differences across degree programs are also evident: L021 is associated with a higher risk of failure and dropout compared to L025, while L026 shows a higher likelihood of inactivity-related withdrawal. These findings highlight the importance of institutional and curricular structures in shaping student trajectories.

From a methodological perspective, the comparison between multinomial logistic regression and the Fine–Gray competing risks model reveals complementary advantages [10,16]. The multinomial framework provides a global view of academic outcomes and achieves strong predictive performance when full academic trajectories are used, as reflected by validation metrics such as the Brier Score. However, it offers limited insight into early-stage dropout dynamics. In contrast, the Fine–Gray model explicitly accounts for the temporal structure of the data and the presence of competing risks, providing a more appropriate framework for analysing student progression over time. Overall, there is a clear trade-off between predictive accuracy and temporal interpretability: full-trajectory models yield higher accuracy, whereas early-stage models are more suitable for timely identification of at-risk students.

The competing risks analysis further supports these findings. Higher academic performance increases the probability of graduation, whereas a greater accumulation of remaining credits is associated with delayed completion or dropout. Pre-university performance has a positive effect on academic success and reduces the risk of dropout and inactivity; however, its influence diminishes over time. In particular, strong early academic performance acts as a protective factor against attrition [8,12], underscoring the importance of early academic integration.

From a policy perspective, the results suggest that interventions based exclusively on pre-university profiles are insufficient for supporting long-term student success. Continuous monitoring of early academic performance provides a more reliable basis for identifying at-risk students and designing targeted interventions such as tutoring, mentoring, and personalised academic support. The observed heterogeneity across degree programmes further indicates that support strategies should be tailored to specific curricular contexts rather than applied uniformly.

The comparison of cumulative incidence functions between training and validation samples provides evidence of good model calibration in the early follow-up period, while some discrepancies emerge in later years, particularly for dropout-related outcomes. This graphical assessment complements the regression results by supporting the evaluation of model stability and generalisability. Consistently, time-dependent AUC values indicate strong discriminative performance of the Fine–Gray model, particularly for graduation and dropout at later time points, with gradual improvement for inactivity-related outcomes.

Finally, the distribution of linear predictors shows clear separation between outcome groups for graduation and inactivity-related withdrawal, while greater overlap is observed between dropout and inactivity. This suggests that dropout trajectories share characteristics with multiple academic pathways, making them more difficult to distinguish. Taken together, AUC measures [4], predictor distributions, and cumulative incidence comparisons provide a consistent assessment of model performance, indicating strong overall discrimination but also highlighting areas of reduced predictive clarity.

Several limitations should be acknowledged. The analysis relies on administrative data from a single institution and does not include potentially relevant variables such as socio-economic background, motivation, or student engagement. Future research should extend this framework to multi-institutional settings and incorporate richer individual-level information to improve predictive performance and generalisability

References

1. Bean, J.P.: Dropouts and turnover: the synthesis and test of a causal model of student attrition. Res. High. Educ. **12**(2), 155–187 (1980)
2. Bean, J.P.: Interaction effects based on class level in an explanatory model of college student dropout syndrome. Am. Educ. Res. J. **22**(1), 35–64 (1985)

3. Bettinger, E.P., Long, B.T., Oreopoulos, P., Sanbonmatsu, L.: The role of application assistance and information in college decisions: results from the H & R Block FAFSA experiment. Quart. J. Econ. **127**(3), 1205–1242 (2012)
4. Blanche, P., Dartigues, J.F., Jacqmin-Gadda, H.: Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat. Med. **32**(30), 5381–5397 (2013)
5. Blanche, P., Dartigues, J.F., Jacqmin-Gadda, H.: timeROC: Time-Dependent ROC Curve Estimation from Censored Survival Data (2024). https://CRAN.R-project.org/package=timeROC, R package version 4.0.5
6. Bound, J., Lovenheim, M.F., Turner, S.: Why have college completion rates declined? an analysis of changing student preparation and collegiate resources. Am. Econ. J. Appl. Econ. **2**(3), 129–157 (2010)
7. Buenaño, E., Beletanga, M.J., Mancheno, M.: What factors are relevant to understanding dropout? analysis at a Co-Financed University in Ecuador and Policy Impli cations, Using Survival Cox Models. J. Latinos Educ. **23**(4), 1400–1415 (2024). https://doi.org/10.1080/15348431.2023.2271570
8. DesJardins, S.L., Ahlburg, D.A., McCall, B.P.: A temporal investigation of factors related to timely degree completion. J. Higher Educ. **73**(5), 555–581 (2002)
9. Fierro Saltos, W.R., Fierro Saltos, F.E., Elizabeth Alexandra, V.S., Rivera Guzmán, E.F.: Leveraging artificial intelligence for sustainable tutoring and dropout prevention in higher education: a scoping review on digital transformation. Information **16**(9), 819 (2025)
10. Fine, J.P., Gray, R.J.: A proportional hazards model for the subdistribution of a competing risk. J. Am. Stat. Assoc. **94**(446), 496–509 (1999)
11. Heisey, D.M., Patterson, B.R.: A review of methods to estimate cause-specific mortality in presence of competing risks. J. Wildl. Manag. **70**(6), 1544–1555 (2006)
12. Herzog, S.: Measuring determinants of student return vs. dropout/stopout vs. transfer: a first-to-second year analysis of new freshmen. Res. Theory & Pract. **6**(1), 57–74 (2005)
13. Hosmer, D.W., Lemeshow, S.: Goodness of fit tests for the multiple logistic regression model. Commun. Stat. Theory and Methods **9**(10), 1043–1069 (1980)
14. Leow, T., Li, W.W., Miller, D.J., McDermott, B.: Prevalence of university non-continuation and mental health conditions, and effect of mental health conditions on non-continuation: a systematic review and meta-analysis. J. Ment. Heal. **34**(2), 222–237 (2025). https://doi.org/10.1080/09638237.2024.2332812, pMID: 38588717
15. Pusztai, G., Fényes, H., Kovács, K.: Factors influencing the chance of dropout or being at risk of dropout in higher education. Educ. Sci. **12**(11) (2022). https://doi.org/10.3390/educsci12110804, https://www.mdpi.com/2227-7102/12/11/804
16. Putter, H., Schumacher, M., van Houwelingen, H.C.: On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: the Fine-Gray model revisited. Biom. J. **62**(3), 790–807 (2020)
17. SAS: SAS/STAT® 9.22 User's Guide. https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#logistic_toc.htm. Accessed 27 Nov 2025
18. SAS: SAS/STAT® 9.22 User's Guide. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#phreg_toc.htm. Accessed 26 Feb 2026
19. Tinto, V.: Dropout from higher education: a theoretical synthesis of recent research. Rev. Educ. Res. **45**(1), 89–125 (1975). https://doi.org/10.3102/00346543045001089

20. Tinto, V.: Leaving College: Rethinking the Causes and Cures of Student Attrition. University of Chicago Press, Chicago (1993)
21. Vaarma, M., Li, H.: Predicting student dropouts with machine learning: an empirical study in finnish higher education. Technol. Soc. **76**, 102474 (2024). https://doi.org/10.1016/j.techsoc.2024.102474, https://www.sciencedirect.com/science/article/pii/S0160791X24000228