International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 24, Number 1

February – 2023

Using Survival Analysis to Identify Populations of Learners at Risk of Withdrawal: Conceptualization and Impact of Demographics

Juan Antonio Martínez-Carrascal[1], Martin Hlosta[2], and Teresa Sancho-Vinuesa[1]
[1]Universitat Oberta de Catalunya; [2]Institute for Research in Open, Distance and eLearning, Swiss Distance University of Applied Sciences

Abstract

High dropout rates constitute a major concern for higher education institutions, due to their economic and academic impact. The problem is particularly relevant for institutions offering online courses, where withdrawal ratios are reported to be higher. Both the impact and these high rates motivate the implementation of interventions oriented to reduce course withdrawal and overall institutional dropout. In this paper, we address the identification of populations of learners at risk of withdrawing from higher education online courses. This identification is oriented to design interventions and is carried out using survival analysis. We demonstrate that the method's longitudinal approach is particularly suited for this purpose and provides a clear view of risk differences among learner populations. Additionally, the method quantifies the impact of underlying factors, either alone or in combination. Our practical implementation used an open dataset provided by The Open University. It includes data from more than 30,000 students enrolled in different courses. We conclude that low-income students and those who report a disability comprise risk groups and are thus feasible intervention targets. The survival curves also reveal differences among courses and show the detrimental effect of early dropout on low-income students, worsened throughout the course for disabled students. Intervention strategies are proposed as a result of these findings. Extending the entire refund period and giving greater academic support to students who report disability are two proposed strategies for reducing course withdrawal.

Keywords: course withdrawal, demographics, distance education and online learning, dropout, intervention design, survival analysis

Using Survival Analysis to Identify Populations of Learners at Risk of Withdrawal: Conceptualization and Impact of Demographics

Academic withdrawal constitutes one of the biggest challenges in education, in particular for online higher education (OHE) institutions, where withdrawal ratios are reported to be higher (Bawa, 2016; Simpson, 2010). Aside from its macroeconomic impact, withdrawal causes frustration in terms of expectations, as well as being a waste of time and money from the student's perspective (Lee & Choi, 2013; Simpson, 2010). These facts justify the interest of and motivate these institutions in designing targeted interventions aimed to reduce it.

A critical first step towards a successful intervention is the accurate and reliable identification of learners at risk (Rienties et al., 2016). This identification is mostly understood in terms of prediction. Most research works focus on determining individual risk and on increasing prediction ratios rather than on understanding the reasons behind the risk. While determining if a particular student is at risk can be valuable, the essential issue when considering an intervention is identifying a common risk factor behind a group of learners who may constitute an intervention target.

Furthermore, timely execution is essential. Time plays a particularly relevant role when designing and implementing interventions oriented to reduce course withdrawal and overall university dropout. The moment when a student decides to abandon a course is critical in terms of the intervention design. At the course level, Simpson (2010) showed that 40% of new students at the Open University withdraw from courses before the first assignment. At the university level, Grau-Valldosera et al. (2019) showed that periods of non-enrolment could result in dropout, despite the intention to continue at the time of the break. In both cases, it would be inefficient to implement interventions after the student has effectively dropped out.

When added to the relevance of time, the concept of population at risk—rather than individual at risk—makes us consider survival analysis as a suitable technique. However, a literature review revealed that research using this technique mainly focused on analysing university dropout (Cobre et al., 2019) or attrition in MOOCs (Rizvi et al., 2022; Xing et al., 2019) and was not linked to interventions. Our article focuses on the use of survival analysis as part of the intervention process, detecting populations of learners at risk of withdrawal at the course level in regular OHE courses. The method described will determine the significance and influence of a set of variables on course withdrawal, providing information to select intervention targets and coherent strategies. Additionally, survival curves will provide additional insight which will help the intervention design.

Besides setting the conceptual framework, we performed a practical implementation based on an open dataset from a world-leading online university: The Open University Learning Analytics Dataset (OULAD; Kuzilek et al., 2017). This dataset contains data from more than 30,000 students enrolled in 22 online course editions from different disciplines, including the withdrawal date for students who abandon different courses. Based on these data, we analysed the impact of students' demographics on withdrawal, determining risk factors and quantifying their impact. Demographics have been identified as some of the causes behind withdrawal (Hachey et al., 2022; Muljana & Luo, 2019) and constitute key features for early

dropout prediction in online environments (Radovanovic et al., 2021). Nonetheless, the proposed method is applicable to any other variable of interest such as academic performance, background, or psychological features/traits that may impact it.

Literature Review

The Concept of Withdrawal

The analysis of dropout has long been present in educational literature, with 1900–1950 being considered the age of early development, broadening horizons in the 1990s, and showing rising interest in recent years. Compilations can be found linked to higher education (Aljohani, 2016; Behr et al., 2020; Larsen et al., 2013) and specifically to online scenarios (Hachey et al., 2022; Lee & Choi, 2011; Muljana & Luo, 2019; Xavier & Meneses, 2020).

Works by Tinto (1975) expounded upon one of the most relevant initial models explaining dropout in traditional education. The core of this theory is the student integration model, where persistence is explained by a student's motivation and ability to match the social and academic characteristics of the institution where she is studying. Years later, Bean (1985) introduced the student attrition model, which relies on the concept of behavioural intention, where dropout is conditioned by a mixture of academic, social-psychological, environmental, and socialisation factors.

These two theories and their combination in Cabrera et al. (1992) and later in Rovai (2003) have been at the core of subsequent studies on the topic. According to Rovai (2003), academic performance and dropout are a combination of student characteristics, student skills, external factors, and internal factors. These four make up the composite persistence model (CPM) and reflect the multivariate nature of dropout.

The term *university dropout* is commonly used to describe situations where students leave the university before obtaining a formal degree (Larsen et al., 2013). Behind this definition lies a complex phenomenon, evidenced by the list of related terms such as dropout, departure, withdrawal, failure, non-continuance or non-completion (Xavier & Meneses, 2020). Dropout is the opposite of retention, defined as "continued student participation in a learning event to completion, which in higher education is a course, program, institution, or system" (Berge & Huang, 2004, p. 3).

At the course level, most papers dealing with withdrawal do not provide a formal definition (77.78% according to a recent scoping review; Xavier & Meneses, 2020). In our research, we used the definition provided by the Open University as "cease studying a module without the intention to resume the study of that module" (Open University, 2022, p. 6).

Approaches for the Identification of Populations at Risk: Survival Analysis

The first stage of a correct intervention design is an accurate and reliable identification of learners at risk (Rienties et al., 2016). Surveys and different data mining techniques are typical approaches used in this

identification. Prevalent techniques include decision trees and random forest (Behr et al., 2020), but a whole set of methods can be found in the literature (Xing et al., 2019). However, only a low percentage of studies make use of longitudinal data approaches and, in particular, survival analysis. Ameri et al. (2016) indicated that "there is only a limited attempt at using these methods in student retention problems" (p. 904). Xing et al. (2016) also showed that the performance of classical techniques used to predict dropout could be improved by accommodating temporal modelling approaches.

The use of survival analysis at the course level in the literature is focused on MOOC scenarios (recently Moreno-Marcos et al., 2019; Rizvi et al., 2022; Xing et al., 2019). The existing studies covering survival analysis in OHE all focus on analysing the semesters when students drop out from the university rather than withdrawal from within courses (Ameri et al., 2016; Cobre et al., 2019; Villano et al., 2018). Two of the studies (Ameri et al., 2016; Villano et al., 2018) focused more on comparing the prediction capability of survival methods to existing techniques. On the other hand, Cobre et al. (2019) tried to identify in which semesters students are most likely to drop out, applied in two different academic programmes in Brazil.

Although some studies (Ameri et al., 2016; Villano et al., 2018) highlighted its interpretation of results and its suitability for analysing underlying student issues and helping the design of interventions, none of the studies examined survival analysis itself. Moreover, to the best of our knowledge, none of the studies examined within-course withdrawal. Considering the importance of the moment of withdrawal as well as the method's longitudinal approach and interpretability, we consider it a suitable approach to designing targeted actions oriented to reducing withdrawal.

Influence of Demographics

Rovai's model indicates the relevance of a student's personal factors linked to dropout in online studies. Focusing on online education, different compilations (Hachey et al., 2022; Lee & Choi, 2011; Muljana & Luo, 2019) investigated the relevance of these factors and showed a lack of consensus among the studies analysed. As noted by Lee and Choi (2011), "findings of many studies were incompatible with one another regarding the relationship between demographics and online students' persistence in online courses" (p. 603).

In particular, the correlation between gender and course withdrawal is unclear. Some works have indicated a relation, which can even depend on the field of study (Cochran et al., 2014). This work indicated that males showed higher withdrawal rates in courses linked to disciplines such as education or health, but lower in those related to business and math. A large number of studies, however, did not establish a correlation between gender and withdrawal (James et al., 2016; Strang, 2017).

Regarding age, OHE students are older than those in face-to-face learning environments. Once enrolled, older students would have a lower dropout rate (James et al., 2016). Other research, however, did not identify any age-related effects (Strang, 2017).

Prior academic achievement is linked to persistence in online learning (Lee & Choi, 2011) and can even be used for prediction (Hachey et al., 2014). Regarding socioeconomic status, it is considered a relevant factor

(Hachey et al., 2022). When considering re-enrolment, having a full-time job and cost factors have a negative impact on retention (Grau-Valldosera et al., 2019). Specifically, students requiring financial aid to re-enrol show higher dropout (Cochran et al., 2014).

Few references can be found to the impact of disability. However, in a few studies, disability is cited by some students as a reason for withdrawal (Shah & Cheng, 2019).

Although several research works have used the OULAD dataset, none has been found covering demographics' role in withdrawal. The closest analysis found (Rizvi et al., 2019) considered the impact of these factors on academic outcomes in terms of pass-fail. This study reported that region, neighbourhood poverty level, and prior education constitute strong predictors of failure.

Research Questions

Considering the lack of studies that analyse withdrawal at the course level in regular OHE with a longitudinal approach, the relevance of reliable identification of learners at risk, and the potential of survival analysis, we formulated this research question:

RQ1: How can survival analysis be used to identify populations of learners at risk of withdrawal at the course level, providing insight into the factors behind that withdrawal?

Additionally, considering both the relevance of time and the potential impact of demographics on withdrawal, we posed a second research question, addressing practical implementation:

RQ2: What is the specific impact of demographic factors over time on course withdrawal? Which of these factors impact the withdrawal regardless of the course itself?

Specifically, we decided to analyse the impact of these demographic characteristics based on the OULAD dataset: (a) age, (b) gender, (c) disability, (d) region, (e) previous academic background, and (f) student's economic situation.

As mentioned, the OULAD dataset includes data from 22 editions of 6 different courses. Detailed information on the dataset is provided in the next section.

Method

Survival Analysis

Survival analysis is "a collection of statistical procedures for data analysis where the outcome variable of interest is time until an event occurs" (Clark et al., 2003a, p. 237). The method is particularly used in medical research, where survival time or time to relapse is under consideration (Bradburn et al., 2003a, 2003b; Clark et al., 2003a, 2003b). The portability of the method to other disciplines has been suggested in recent studies (Emmert-Streib & Dehmer, 2019).

Kaplan-Meier (KM) estimates and, specifically, KM curves are common in most survival analyses when the goal is to compare two populations. They are the simplest way to compute survival over time (Clark et al., 2003a). KM estimates help to establish whether life expectancy is different for different populations who have different characteristics, or whether a specific treatment can be more advisable than others. Linked to this estimation, the hazard function indicates the probability of not surviving beyond a certain point in time.

The statistical significance of the resulting curves can be checked with the log-rank test (Clark et al., 2003a). This test compares the estimates of the hazard functions of the two groups at each observed event time under the null hypothesis that both groups share the same hazard functions. The original test assigns equal weight to early and late events. Modified versions use weighted functions. In particular, Peto-Peto's log-rank test (Peto & Peto, 1972) assigns weights depending on the estimated percentile of the failure time distribution, giving higher weight to earlier events, and is commonly used within this group.

However, KM estimates cannot quantify the impact of a given parameter, particularly when dealing with different variables, i.e., the covariates. When this is required, parametric methods must be used. Fully parametric methods need to assume statistical distribution in the data. If this distribution is known, they can provide more precise models. Semi-parametric methods have the advantage of being able to quantify the impact without assuming a specific distribution. The most used semi-parametric method is the Cox proportional hazards model (Bradburn et al., 2003b). This model is based on a proportional hazard assumption and computes a baseline time-dependent hazard associated with a reference group. This hazard is modified based on the multiplicative effect of the values of the different covariates, whose individual influence is considered constant over time. Once the method is computed, the assumptions need to be checked.

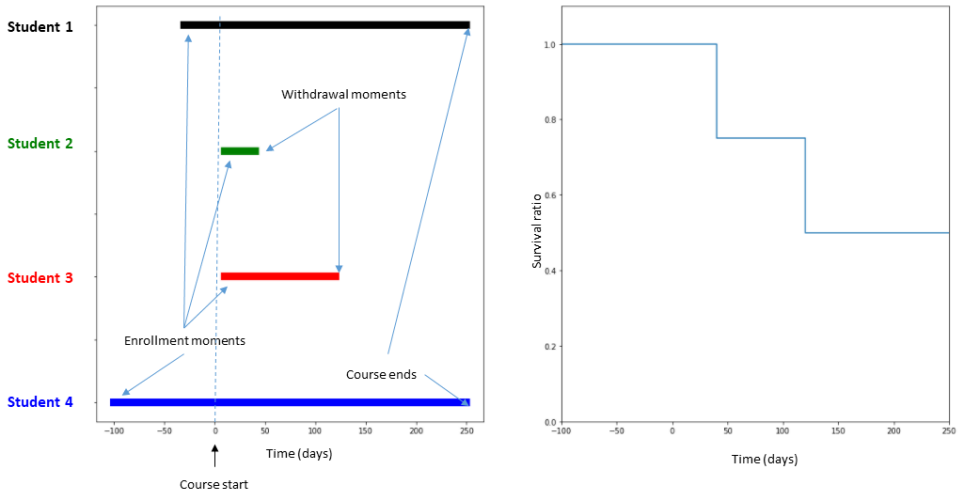
Porting Survival Analysis to Withdrawal Analysis

Approaching a generic problem through survival analysis requires a precise mapping of three concepts: the lifespan, the event under consideration, and the period of observation (Clark et al., 2003a).

In the case of withdrawal, the number of days a student remains enrolled after the specific course starts constitutes the lifespan. The event under consideration is the withdrawal decision. The analysis would also need to monitor on a periodical basis whether the student has withdrawn. To set up a common reference among courses, the course start date would be considered as t = 0. Negative values indicate days before the course starts. Survival curves reflect how a population survives after a certain time. Figure 1 depicts these concepts in a hypothetical course lasting 250 days with four students enrolled.

Figure 1

Graphical View of a Hypothetical Course and Associated Survival Curve



Note. Left panel: Enrolment and withdrawal or completion dates for four students. Right panel: Associated survival curve for this group.

On the left, Figure 1 shows four students enrolling on different dates. The first is Student 4, enrolling 100 days before the course starts. Students 2 and 3 enrol ten days after the course has started. In this example, Student 2 withdraws shortly after enrolment (40 days after the course starts), while Student 3 withdraws 120 days after the course starts. Students 1 and 4 complete the course. The associated survival curve for this group is shown on the right, where we can see that the final survival ratio is 0.5 (2 out of 4 students). The curve provides not only the final ratio but a graphical view of its evolution.

KM plots provide a graphical view of the individual impact of specific covariates. To aggregate and quantify the impact of those found relevant, we used the Cox proportional hazards model, due to its simplicity compared to parametric methods.

Dataset

These concepts were translated into practice using the public dataset offered by The Open University (OU; Kuzilek et al., 2017). This dataset provides information about 22 editions (*presentations* in the dataset nomenclature). A total of 32,593 students are enrolled in these courses. The typical presentation length is around nine months.

Courses included in the dataset were offered via a virtual learning environment (VLE), and each had over 500 students. While part of the OU course portfolio, students without a previous academic background could also enrol. Table 1 summarises a high-level view of enrolment and academic results in the courses included. Academic results are summarised in four categories: withdraw, fail, pass, or distinction.

Table 1

Global View of Enrolment and Academic Results

Indicator	Enrolled	Withdraw	Fail	Pass	Distinction
Number of students	32,593	10,156	7,052	12,361	3,024
Percentage of total (%)	100	31.16	21.64	37.93	9.28

As Table 1 shows, withdrawal constituted 31.16% of the global population enrolled. The distribution of academic results was not homogenous among courses as displayed in Table 2.

Table 2

Enrolment and Academic Results (Per-Course View)

Course	Students	Withdraw	Fail	Pass	Distinction
	n	%	%	%	%
AAA	747	16.73	12.18	65.19	5.89
BBB	7,903	30.18	22.32	38.93	8.57
CCC	4,434	44.54	17.61	26.61	11.23
DDD	6,266	35.86	22.49	35.54	6.11
EEE	2,934	24.61	19.15	44.10	12.13
FFF	7,758	30.96	22.02	38.39	8.64
GGG	2,534	11.52	28.73	44.12	15.63

Note. Courses are identified with anonymised course names (i.e. AAA) in the OULAD dataset.

These high withdrawal ratios may be explained by the fact that they constitute regular OU courses, with high academic standards, but at the same time, require no prior qualification for enrolment. All courses share a common framework for evaluation, including a set of tutor-marked assignments and optionally some computer-marked assignments. Also, there is usually a final exam at the end of each course.

With respect to those students withdrawing, the dataset includes information regarding the date of withdrawal. This date is either the date on which the student notified the university of her withdrawal or the date on which the student's participation in the module ceased, whichever came first. The Open University actively seeks to reduce withdrawal and may monitor online student activity to detect it. Students considering withdrawal are advised to contact the module instructor and, if their decision is final, formally report their decision (Open University, 2022).

The dataset also includes some personal information. Table 3 summarises those characteristics in the dataset considered relevant to our study.

Table 3

Characteristics in the OU Dataset Linked to the Research Questions

Scope	Variable	Meaning
Presentation	length	Length in days of the module presentation
	date_registration	The day the student registers for the module presentation
Registration	date_unregistration	The day the student unregisters from the module presentation
Demographic characteristic	gender	Gender of the student (male/female)
	region	The geographic region, where the student lived while taking the presentation
	imd_band	The index of multiple deprivation (IMD) band of the place where the student lived during the module presentation
	highest_education	The highest student education level on entry to the module presentation (5 bands)
	age_band	Age band of the student (3 bands)
	disability	Indicates whether the student has declared a disability

Note. Variable names used match those in the OULAD dataset.

This information was required to approach RQ2. Age, gender, and disability are available directly in the dataset. Previous academic background is expressed as the highest educational level the student achieved before the module started. Region indicates the area where the student lives. Student economic situation is expressed by the index of multiple deprivation (IMD) used in the UK (Kuzilek et al., 2017; Rizvi et al., 2019). The dataset presents IMD figures in bands ranging from 0%-10% to 90%-100%; 0%-10% means that a student lives in the most deprived UK areas, while 90%-100% points to the least deprived areas.

Results

The preceding section identifies two main steps for practical implementation:

1. Use KM estimates to determine populations at risk and the impact of individual covariates on withdrawal.

2. Analyse the combined impact, quantifying the simultaneous effect through Cox proportional hazards model.

The Cox model requires a prior setup of reference values for the covariates. For the categorical variables shown in Table 3, we generated dummy variables and considered the values shown in Table 4 as reference values.

Table 4

Reference Groups for the Computation of the Cox Model

Covariate	Reference group
Gender	Female
Region	North region
Highest education	A level or equivalent
IMD band	50%–60%
Age band	Under 35
Disability	No

Note. IMD = index of multiple deprivation.

When the number of possible values was high, we selected reference values that reflected a more central position (e.g., IMD band = 50%–60%). For the specific IMD scales, we grouped low IMD scales (0%–30%) and high IMD scales (above 80%) to reduce the overall number of values.

Significant Differences Based on IMD Band, Prior Education, and Declared Disability

Covariates to perform KM estimates were extracted from Table 3. Using Peto-Peto log-rank tests, we computed *p*-values. Data in Table 2 reflect that different courses show differences in withdrawal ratios. For this reason, we also performed a per-course analysis to determine whether covariates were significant both at the global and individual course levels. The results are shown in Table 5.

Table 5

Statistical Significance of Covariates at the Global and Individual Course Levels

Covariate	Global	Individual course						
		AAA	BBB	CCC	DDD	EEE	FFF	GGG
Gender	ns	*	ns	ns	***	*	*	ns
Region	****	ns	****	ns	ns	**	ns	ns
Highest education	****	ns	****	****	****	***	****	ns
IMD band	****	ns	****	****	****	**	****	ns
Age band	****	ns	**	**	ns	ns	ns	ns
Disability	****	ns	ns	***	****	ns	****	*

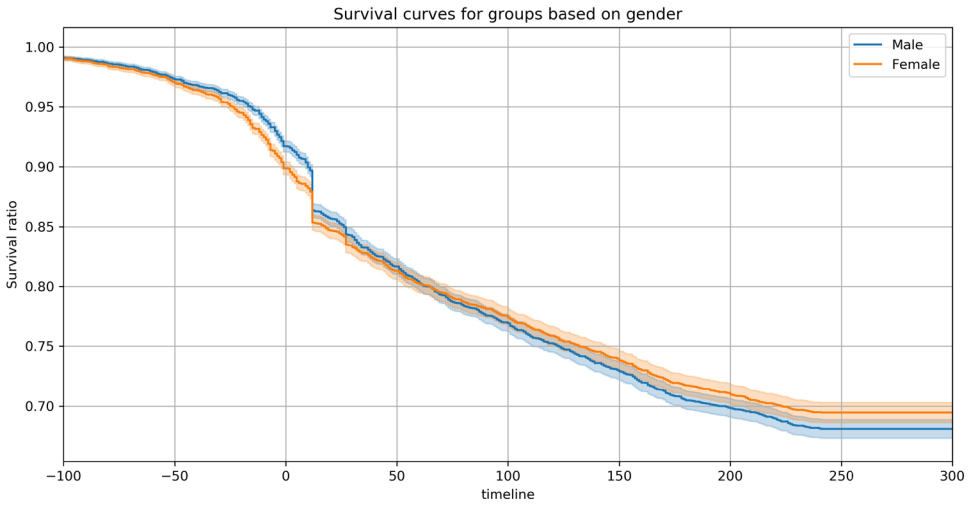
Note: ns = non-significant. IMD = index of multiple deprivation. Courses are identified with anonymised course names (i.e. AAA) in the OULAD dataset. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$. **** $p < 0.0001$.

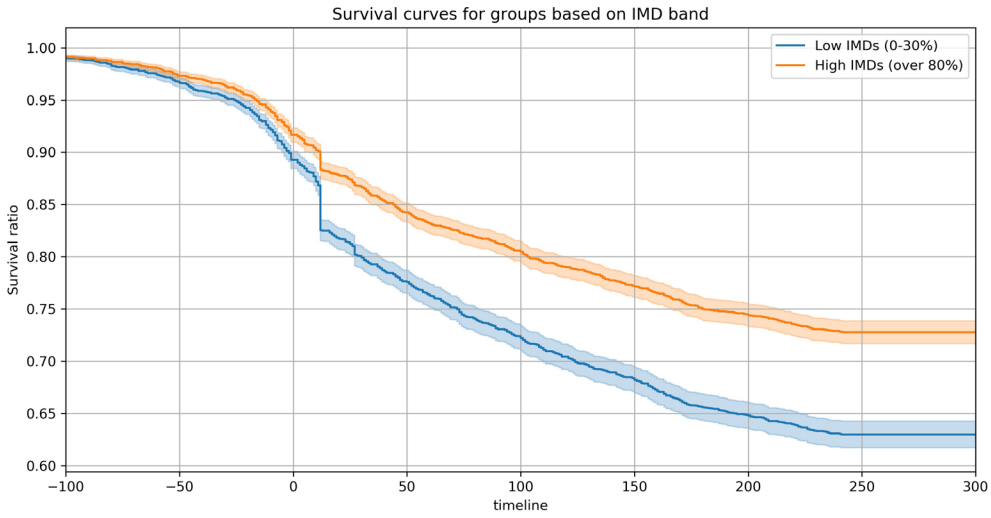
Prior highest education level and IMD band have a clear impact when considering either the global data set or individual courses. At the course level, more data would be needed for course AAA to provide statistically significant results. While age looks relevant globally, its effect disappears in most courses when analysed individually. Thus, no definite conclusion can be extracted at this stage. More data would also be needed, as there is a low ratio of students in one of the scales considered in the dataset.

KM plots help to visualise differences. As an example, we show the global impact of two covariates: gender—not significant according to the test—and the IMD band, which is significant. For clarity, in the case of IMD plots, we compared the high (> 80%) and low (< 30%) groups. The results are shown in Figure 2.

Figure 2

Survival Curves for Different Groups Based on Gender and IMD Band





Note. Top panel: The survival curve for gender. Bottom panel: The survival curve for IMD band. IMD = index of multiple deprivation.

Figure 2 shows minor differences based on gender. Regarding IMD bands, this figure reflects higher withdrawal ratios for the low IMD group, with a higher impact of early withdrawal.

Previous Risk Factors also Present when Considering Simultaneous Effect

We used the Cox model to evaluate and quantify the simultaneous effect of the different covariates. The final Cox models were developed with two strata variables (course and disability) and a set of dummy variables linked to IMD band, region, gender, and previous higher education. Table 6 summarises those variables that appear relevant at either the global or individual course level.

Table 6

Hazard Risk Factor Relative to the Reference Group Based on the Values of Covariates

Covariate	Individual course							Global
	AAA	BBB	CCC	DDD	EEE	FFF	GGG	
Gender: Male	ns	ns	ns	0.83	ns	0.90	ns	0.89
Region: East Midlands	ns	ns	1.23	ns	ns	ns	ns	1.14
Region: London	ns	ns	ns	ns	1.41	1.20	ns	ns
Region: West Midlands	ns	ns	ns	ns	1.39	ns	ns	1.12
Highest education: HE qualification	ns	ns	0.83	ns	ns	ns	ns	0.93
Highest education: Lower than A Level	ns	1.41	1.41	1.30	1.48	1.42	ns	1.38

Highest education: No formal qualifications	ns	1.73	1.48	1.72	2.38	1.38	ns	1.63
Highest education: Post-graduate qualification	ns	ns	0.56	ns	ns	ns	5.51	0.75
IMD band: 0%–30%	ns	1.18	1.33	1.37	1.56	1.44	ns	1.35
IMD band: 30%–40%	ns	ns	ns	1.25	ns	ns	ns	1.14
IMD band: 40%–50%	ns	ns	1.21	1.35	1.75	ns	ns	1.21
IMD band: 80%–100%	ns	ns	ns	ns	ns	ns	1.69	ns

Note: Only covariates significant in at least one course shown. IMD = index of multiple deprivation.

To understand the impact of disability, we compared baseline survival functions for the different strata generated at the course level. Table 7 reflects the withdrawal increase ratio for individual courses when disability was a factor.

Table 7

Impact of Disability on Withdrawal Risk—Individual Course Level

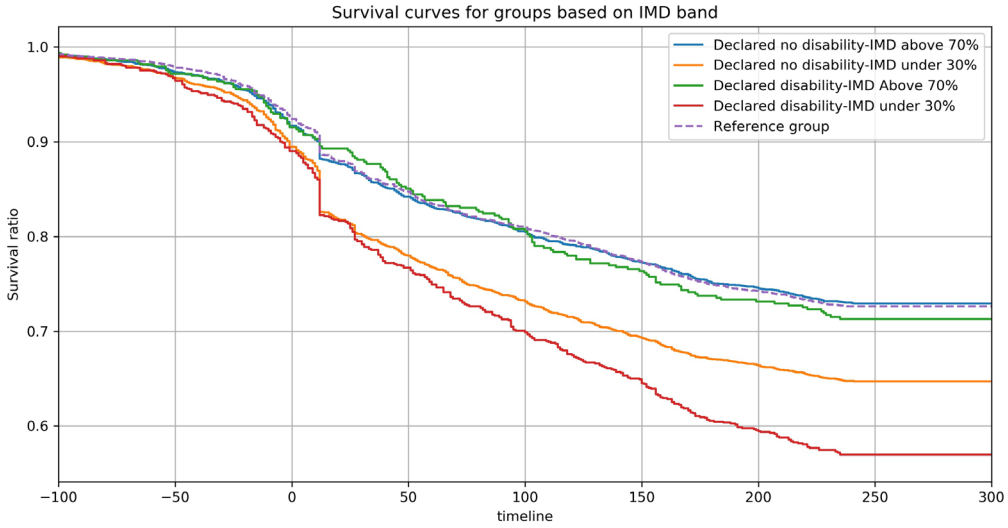
Indicator	Individual course					
	BBB	CCC	DDD	EEE	FFF	GGG
Withdrawal risk increase (declared disability vs declared no disability)	1.14	1.19	1.49	0.99	1.43	1.45

Note. Courses are identified with anonymised course names (i.e. BBB) in the OULAD dataset. Course AAA showed inconclusive results and was omitted from this table.

As a final check, we performed a graphical comparison of withdrawal differences based on the findings above. We generated populations based on the combination of IMD differences—high versus low group—and disability. The results are shown in Figure 3, where a reference group based on data in Table 4 (no disability, IMD band 50%–60%) is also reflected.

Figure 3

Survival Curves for Groups With and Without Declared Disability in High and Low IMD Bands



Note: IMD = index of multiple deprivation.

Figure 3 clearly shows withdrawal risk differences among groups. Besides final survival expectancy—with differences around 35.88% by the end of the course—low-income students drop out earlier. Also, the multiplicative effect of disability and low IMD is clearly displayed. Being in the high IMD group does not significantly reduce withdrawal rates when compared to the reference group. The impact of these findings on potential intervention designs will be addressed in the next section.

Discussion

Two research questions were addressed in this work. The first one, regarding the use of survival analysis, aimed to detect populations at risk of withdrawal and the factors behind it. The second one aimed to translate these concepts into practice, determining the relevance and impact of demographics.

From a methodological perspective, the basics behind the answer to RQ1 are covered in the subsection covering the portability of survival analysis to learning analytics scenarios. Identifying students at risk of withdrawal through survival analysis has required the mapping of three concepts: the event under consideration (the withdrawal decision), the period of observation (a course), and the lifespan (the time the student remains in the course). This mapping allows us to identify both at-risk populations and the associated risk factors. Figure 1 concentrates on the basics behind this mapping.

Survival curves provide a graphical insight into the differences among populations based on a set of factors (Figures 2 and 3 offer clear examples). These curves constitute a relevant difference from other data mining techniques. They do not only provide information on final withdrawal ratios, but also show when withdrawal occurs. Statistical relevance of a given factor can also be determined (see Table 5 for examples), and for those factors considered relevant, the impact can also be quantified (Tables 6 and 7 serve as examples for this point). These facts make survival analysis a particularly suitable technique for analysing withdrawal.

All in all, figures 1 (from a theoretical perspective), 2, and 3 (from a practical approach), combined with the data in tables 6 and 7, demonstrate the potential of survival analysis identifying populations of learners at risk.

The second question (RQ2) translates methodology into practice. The application of the method indicates that certain demographic characteristics have an impact on course withdrawal and that this impact is dependent on the course itself. Specifically, three analysed factors increase withdrawal risk: a previous level of education below the reference group (A level), a low IMD band, and a declared disability (see Table 5). As mentioned, the specific impact is different depending on the course (see Table 7). We can compare these findings with previous literature regarding the impact of demographics on withdrawal.

Withdrawal and Demographics: Comparison with the Literature

From a global perspective, the influence of these personal background factors is consistent with the theoretical models (Bean, 1985; Cabrera et al., 1992; Rovai, 2003; Tinto, 1975) and justifies the interest that literature compilations put on them, in particular in OHE (recently, Hachey et al., 2022; Muljana & Luo, 2019).

Our results have shown a different impact for age and gender across the analysed courses, supporting the inconclusive findings reflected in Hachey et al., 2022 and Muljana and Luo, 2019. We have not found that being male reduces risk in some courses, while increasing it in others (as found in Cochran et al., 2014). However, we agree that the relevance and the specific impact of a factor depend on the course under analysis.

Regarding the economic situation, our results at course level are aligned with those indicating the impact of financial hardship at course and university level (Cochran et al., 2014; Grau-Valldosera et al., 2019). Our work indicates a direct relationship between socioeconomic inequality and educational disadvantage, as shown in the lower panel in Figure 2.

The impact of a poor academic background on withdrawal is consistent with earlier research linking lower previous achievement to higher university dropout (Cochran et al., 2014; Lee & Choi, 2011).

Disability is one of the potential reasons behind some dropouts according to Shah and Cheng (2019). Our results confirm this fact and quantify its impact on course withdrawal. Our findings indicate that students with disabilities taking online courses would be more likely to withdraw from these courses, in particular

those students in low IMD bands. Due to our concerns about equity, we believe more studies on this topic should take specific care of anonymisation and ethical issues.

Implications

The intersectionality and the reliable estimation of risk allow us to identify two points for a potential intervention with targeted populations. First, less affluent students could be contacted, even before the start of the course, and offered options regarding financing. Second, disabled students coming from more deprived areas might benefit from continuous support, which might reduce the slowly increasing difference in withdrawal rates when compared with non-disabled students reporting the same economic condition.

While these demographic factors affect all courses analysed, their impact on dropout is different for each course. This difference needs to be considered when evaluating the outcomes of specific interventions.

We can also find factors that show statistical significance in only some courses. To mention just a couple of examples, gender for course DDD or region for course BBB (see Table 5) warrant investigation. For these cases, we encourage a closer look that considers course-specific details which may explain why.

We also remark on the potential of survival analysis to detect situations that would otherwise remain hidden. Figure 2 (lower panel) and Figure 3 reveal a sudden drop around the second week of the course, particularly affecting low-income students. In fact, this week corresponds to the end of the full-refund period for a given course. A potential intervention aimed at reducing withdrawal would consider extending the period of full refund for low-income students. It is important to highlight that this kind of finding would remain hidden if using techniques which focus only on final ratios and not on temporal evolution.

Our detailed analysis also reveals potential fails in intervention design which do not include a proper identification stage. We can consider for instance prior level of education. It is noticeable that course GGG shows a higher risk of withdrawal for students with a previous higher level of education. While this could be shocking at first glance, this course constitutes a propaedeutic course. Those students who already have this knowledge simply abandon the course. Besides this example, and considering potential interventions, the method reveals that analysis at both a global level and at the level of the individual course is critical to properly identify populations at risk.

Finally, survival analysis provides a clear view of the impact of the different factors analysed. For the case of demographics, IMD band, prior educational level, and declared disability have emerged as the most relevant factors in dropout. It is worth noting that these factors emerge as relevant both at the aggregate level and when considering individual courses.

Conclusion

Survival analysis has proven to be a useful tool to reliably identify populations of learners at risk. The method outlined provides risk quantification, and a clear graphical evolutive view. This view highlights insights that otherwise could remain hidden.

Its use has been particularly suited for the analysis of course withdrawal, due to the relevance of time in dropout. We encourage the use of survival analysis as the first stage in the design of interventions aimed at reducing academic dropout. It can also be of interest in learning analytics scenarios where time plays an important role, such as engagement analysis.

Finally, considering the multivariate nature of withdrawal, we advise expanding this research beyond demographics. While focusing on them has shown the method's potential and provided valuable insight, it also constitutes a limiting factor. We encourage the use of the methodology exposed to address the impact of other aspects, such as previous knowledge, activity reflected in VLEs, or course instructional design. Future work should focus on incorporating these dimensions into the analysis to better understand students' behaviour and improve learning experience and academic performance.

Acknowledgements

We would like to thank The Open University for working on the creation of the dataset, and in particular, Professor Bart Rienties from the Institute of Educational Technology for his readiness to provide additional insight to better understand the dataset. Also, Professor Per Bergamin from the Swiss Distance University of Applied Sciences offered insightful feedback during the review process.

References

Aljohani, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher Education Studies*, *6*(2), 1–18. https://doi.org/10.5539/hes.v6n2p1

Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. In S. Mukhopadhyay & C. X. Zhai (Chairs), *CIKM '16: Proceedings of the International Conference on Information and Knowledge Management* (pp. 903–912). ACM. https://doi.org/10.1145/2983323.2983351

Bawa, P. (2016). Retention in online courses: Exploring issues and solutions—A literature review. *SAGE Open*, *6*(1), 1–11. https://doi.org/10.1177/2158244015621777

Bean, J. P. (1985). Interaction effects based on class level in an explanatory model of college student dropout syndrome. *American Educational Research Journal*, *22*(1), 35–64. https://doi.org/10.3102/00028312022001035

Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: A literature review. *Review of Education*, *8*(2), 614–652. https://doi.org/10.1002/rev3.3202

Berge, Z. L., & Huang, Y.-P. (2004). A model for sustainable student retention: A holistic perspective on the student dropout problem with special attention to e-Learning. *Deosnews*, *13*(5), 1–26.

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003a). Survival analysis part II: Multivariate data analysis—An introduction to concepts and methods. *British Journal of Cancer*, *89*, 431–436. https://doi.org/10.1038/sj.bjc.6601119

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003b). Survival analysis part III: Multivariate data analysis—Choosing a model and assessing its adequacy and fit. *British Journal of Cancer*, *89*, 605–611. https://doi.org/10.1038/sj.bjc.6601120

Cabrera, A. F., Castañeda, M. B., Nora, A., & Hengstler, D. (1992). The convergence between two theories of college persistence. *The Journal of Higher Education*, *63*(2), 143–164. https://doi.org/10.2307/1982157

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003a). Survival analysis part I: Basic concepts and first analyses. *British Journal of Cancer*, *89*, 232–238. https://doi.org/10.1038/sj.bjc.6601118

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003b). Survival analysis part IV: Further concepts and methods in survival analysis. *British Journal of Cancer*, *89*, 781–786. https://doi.org/10.1038/sj.bjc.6601117

Cobre, J., Tortorelli, F. A. C., & de Oliveira, S. C. (2019). Modelling two types of heterogeneity in the analysis of student success. *Journal of Applied Statistics*, *46*(14), 2527–2539. https://doi.org/10.1080/02664763.2019.1601164

Cochran, J. D., Campbell, S. M., Baker, H. M., & Leeds, E. M. (2014). The role of student characteristics in predicting retention in online courses. *Research in Higher Education*, *55*(1), 27–48. https://doi.org/10.1007/s11162-013-9305-8

Emmert-Streib, F., & Dehmer, M. (2019). Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, *1*(3), 1013–1038. https://doi.org/10.3390/make1030058

Grau-Valldosera, J., Minguillón, J., & Blasco-Moreno, A. (2019). Returning after taking a break in online distance higher education: From intention to effective re-enrollment. *Interactive Learning Environments*, *27*(3), 307–323. https://doi.org/10.1080/10494820.2018.1470986

Hachey, A. C., Conway, K. M., Wladis, C., & Karim, S. (2022). Post-secondary online learning in the U.S.: An integrative review of the literature on undergraduate student characteristics. *Journal of Computing in Higher Education*. https://doi.org/10.1007/s12528-022-09319-0

Hachey, A. C., Wladis, C. W., & Conway, K. M. (2014). Do prior online course outcomes provide more information than G.P.A. alone in predicting subsequent online course grades and retention? An observational study at an urban community college. *Computers and Education*, *72*, 59–67. https://doi.org/10.1016/j.compedu.2013.10.012

James, S., Swan, K., & Daston, C. (2016). Retention, progression and the taking of online courses. *Journal of Asynchronous Learning Network*, *20*(2), 75–96. https://doi.org/10.24059/olj.v20i2.780

Kuzilek, J., Hlosta, M., & Zdrahal,Z. (2017). Open University Learning Analytics dataset. *Scientific Data*, *4*(1), 170171. https://doi.org/10.1038/sdata.2017.171

Larsen, M. S., Kornbeck, K. P., Kristensen, R. M., Larsen, M. R., & Sommersel, H. B. (2013). *Dropout phenomena at universities: What is dropout? Why does dropout occur? What can be done by the universities to prevent or reduce it?: A systematic review*. Danish Clearinghouse for Educational Research.

Lee, Y., & Choi, J. (2011). A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development*, *59*(5), 593–618. https://doi.org/10.1007/s11423-010-9177-y

Lee, Y., & Choi, J. (2013). Discriminating factors between completers of and dropouts from online learning courses. *British Journal of Educational Technology*, *44*(2), 328–337. https://doi.org/10.1111/j.1467-8535.2012.01306.x

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Kloos, C. D. (2019). Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies*, *12*(3), 384–401. https://doi.org/10.1109/TLT.2018.2856808

Muljana, P. S., & Luo, T. (2019). Factors contributing to student retention in online learning and recommended strategies for improvement: A systematic literature review. *Journal of Information Technology Education: Research*, *18*, 19–57. https://doi.org/10.28945/4182

Open University, The. (2022). *Changing your study plans policy 2022/23*. https://help.open.ac.uk/documents/policies/changing-your-study-plans/files/214/changing-your-study-plans-policy-2022-23.pdf

Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, *135*(2), 185–207. https://doi.org/10.2307/2344317

Radovanović, S., Delibašić, B., & Suknović, M. (2021). Predicting dropout in online learning environments. *Computer Science and Information Systems*, *18*(3), 957–978. https://doi.org/10.2298/csis200920053r

Rienties, B., Boroowa, A., Cross, S., Farrington-Flint, L., Herodotou, C., Prescott, L., Mayles, K., Olney, T., Toetenel, L., & Woodthorpe, J. (2016). Reviewing three case-studies of learning analytics interventions at the Open University UK. In D. Gašević & G. Lynch (Chairs), *LAK '16: Proceedings of the sixth international conference on learning analytics and knowledge* (pp. 534–535). https://doi.org/10.1145/2883851.2883886

Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning: A decision tree based approach. *Computers and Education*, *137*(2), 32–47. https://doi.org/10.1016/j.compedu.2019.04.001

Rizvi, S., Rienties, B., Rogaten, J., & Kizilcec, R. F. (2022). Beyond one-size-fits-all in MOOCs: Variation in learning design and persistence of learners in different cultural and socioeconomic contexts. *Computers in Human Behavior*, *126*(C), Article 106973. https:/doi.org/10.1016/j.chb.2021.106973

Rovai, A. P. (2003). In search of higher persistence rates in distance education online programs. *Internet and Higher Education*, *6*(1), 1–16. https://doi.org/10.1016/S1096-7516(02)00158-6

Shah, M., & Cheng, M. (2019). Exploring factors impacting student engagement in open access courses. *Open Learning*, *34*(2), 187–202. https://doi.org/10.1080/02680513.2018.1508337

Simpson, O. (2010). *"22%—can we do better?": The CWP retention literature review*. Centre for Widening Participation, Open University UK. https://doi.org/10.13140/RG.2.2.15450.16329

Strang, K. D. (2017). Beyond engagement analytics: Which online mixed-data factors predict student learning outcomes? *Education and Information Technologies*, *22*(3), 917–937. https://doi.org/10.1007/s10639-016-9464-2

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Villano, R., Harrison, S., Lynch, G., & Chen, G. (2018). Linking early alert systems and student retention: A survival analysis approach. *Higher Education*, *76*(4), 903–920. https://doi.org/10.1007/s10734-018-0249-y

Xavier, M., & Meneses, J. (2020). *Dropout in online higher education: A scoping review from 2014 to 2018*. Universitat Oberta de Catalunya. https://doi.org/10.7238/uoc.dropout.factors.2020

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, *58*, 119–129. https://doi.org/10.1016/j.chb.2015.12.007

Xing, W., Tang, H., & Pei, B. (2019). Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of MOOCs. *Internet and Higher Education*, *43*, Article 100690. https://doi.org/10.1016/j.iheduc.2019.100690