Further details on the methodology and the predictive modelling approach

Predicting osteopaths joining the register calculations

1. Based on the data availability, we define a two-step procedure: First, we need to predict the number of students enrolled on an osteopathy course, and second, we need to forecast the number of osteopaths joining the register.

2. To predict the number of students on an osteopathic course, we define the next regression model using the Ordinary Least Squares (OLS):

$$y_i = \alpha + \beta_1 x_i + \beta_1 x_i^2 + \varepsilon_i \qquad (1)$$

3. The dependent variable "y_i" represents the enrollment/ student intake in the academic year"i", where "i" indexes the academic year, and "x" represent each academic year. Before predicting the number of osteopaths joining the register, we need to find the parameter of equation (1) based on the statistics shown in Table 1 that better fit the data exhibited by Figure 1. As a result of this process, we obtain the next baseline model:

$$y_i = -5{,}7829 x_i^2 - 4{,}265 x_i + 1310 + \varepsilon_i \qquad (2)$$

4. Next, we employ our baseline equation (2) to predict the dependent variable"y_i", starting with the period 2023-24.

Predicting osteopaths leaving the register calculations

5. In the first step, we predict the number of leavers according to several scenarios. Starting with a normal scenario, we can predict the number of leavers using a linear regression model (Ordinary Least Squares) based on trends exhibited over time relating osteopaths leaving the register. To predict the number of leavers, we define the next regression model using the Ordinary Least Squares (OLS):

$$y_i = \alpha + \beta_1 x_i + \varepsilon_i \qquad (3)$$

6. The dependent variable "y_i" represents the number of leavers in the academic year "i" and "x" represent each academic year. Before predicting the number of osteopaths leaving the register, we need to find the parameter of equation (3) based on the statistics shown in Table 3. Consequently, and using equation (3) we derive the following baseline equation:

$$y_i = 170.2 + 12.4 x_i + \varepsilon_i \qquad (3)$$

7. We then used the baseline equation (3) to predict the dependent variable"y_i", starting the period 2022-23.

Predictive modelling procedure

8. We use a binary response model regression, more precisely, a Logistic regression method (or Logit model) based on the binary dependent variable under study. Binary response models consider as dependent variable a binary, a dichotomous or a dummy variable (1 or 0). In a binary response model, interest lies primarily in the response probability. Logistic regression predicts whether something is true or false, specifically, the probability that a variable is true or false (Wooldridge, 2012).[1] A typical example is the probability of a student succeeding or not succeeding at university based on the University Selection Test Result, among other variables. This issue has been widely examined in the literature under different contexts/environments, using the prominent model Tinto, (2012)[23]. We rely on this literature to predict the number of registrants leaving the register.

9. To predict the number of osteopaths leaving the register, we define the next logistic model:

$$P(y = 1|x) = P(y = 1|x_1, x_2, \ldots . x_k) \quad (4)$$

   Which can be simplified as

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i \quad (5)$$

10. The dependent variable "y_i" represents the registrant and takes the value of 1 for registrants that left the register (leavers). Those who carry on with the register (retention) take the value 0 of the variable. The variables $x_1, x_2, \text{and } x_3$ represent the explanatory variables, i.e., the demographic variables related to the registrants such as age, years on the register, gender, and reason for leaving the register.

11. We carry out this analysis based on a three-step procedure. First, we examine the relationship between the dependent variable and each independent variable. So far, the explanatory variables are the registrant's age ("Age"), the years an osteopath spends on the record ("Years on the register") and the registrant's gender ("Gender"). Table 1 shows the estimation results considering separately each explanatory variable (Model (1) to (3)) and for the full model (Model (4)). Overall, the estimation results exhibit the expected signs and statistically significant results except for Gender. We observe that "Years on the register" and "Age" impact on the probability of leaving the register. As

[1] Wooldridge, J., (2012) Introductory Econometrics: A Modern Approach. Fifth Edition, South-Western, Cengage Learning.

[2] Tinto, V.,2012. Completing College: rethinking institutional action. Chicago: University of Chicago Press.

[3] Tinto (2012) argues that student success is affected by the degree of academic and social integration; in other words, by the fulfilment of educational guidelines and the ties that are woven between their surroundings and them and the higher education institutional agents.

long as an osteopath spends more years on the register it increases the probability of them leaving the register. Similarly, older osteopaths are most likely to leave the register than younger osteopaths. We can see this in Table 5, worst scenario (scenario 3) in the main report. However, no significant relationship between Gender and the probability of leaving the register was found, indicating that gender is not significant in osteopaths leaving the GOsC register.

Table 1: Logit regressions

	Model (1)	Model (2)	Model (3)	Model (4)
Years on the register	0.105(1.112)***			0.014(0.985)**
Age		0.091(1.094)***		0.094(1.098)***
Gender			0.287(1.332)**	0.076(1.078)
Pseudo R2	0.0648	0.2065	0.004	0.2073

The dependent variable equals one if the registrant leaves the record and zero if the registrant continues on the record. Odds ratios are reported in parentheses ***, **,* represents statistical significance at the 1, 5 and per cent level.

12. In the second step, and after analysing the estimation results from the logistic regressions, we are able to predict the probability of a registrant leaving the GOsC. Table 2 shows the main descriptive statistics of the probability of leaving the register. For example, for Model (4), we observe the mean probability of leaving the register is 45,3%, and the minimum and maximum probability are 1% and 98%, respectively. We will use this Model as it presents the best adjustment to the data, specifically, the highest Pseudo R2.

Table 2: Descriptive statistics

	Min	Max	Mean	P25	P50	P75
Model (1)	0.249	0.815	0.453	0.329	0.436	0.547
Model (2)	0.008	0.983	0.452	0.221	0.412	0.655
Model (3)	0.038	0.417	0.452	0.417	0.417	0.499
Model (4)	0.006	0.985	0.453	0.225	0.407	0.659

This table reports for each model the main descriptive statistics of the probability of leaving the register.