COMPARING SUPERVISED AND SEMI-SUPERVISED MACHINE LEARNING APPROACHES IN NTCP MODELING TO PREDICT COMPLICATIONS IN HEAD AND NECK CANCER (HNC) PATIENTS

by

Isa Spiero

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Graduate School of Life Science
Utrecht University

June 2022

Under the supervision of
A.M. (Tuur) Leeuwenberg PhD
Dr. ir. E. (Ewoud) Schuit

Abstract

Head and neck cancer patients (HNC) treated with radiotherapy often suffer from radiation-induced toxicities, most often xerostomia (dry mouth) and dysphagia (difficulty swallowing). As to reduce the risk of toxicities in these patients, Normal Tissue Complication Probability (NTCP) modeling is used to determine the probability to develop toxicities based on patient and treatment characteristics. Currently, most often supervised logistic regression methods are used in NTCP modeling. However, as the toxicity outcomes that are used in NTCP modeling of HNC are often recorded long after treatment started, 'unlabeled' data are also available. In these data the patient and treatment characteristics are recorded, but not yet the toxicity outcomes. Semi-supervised methods are able to incorporate unlabeled data in model development and may thereby gain in predictive performance compared to the current supervised logistic regression models. Here, it will be evaluated how current regression models compare to the semi-supervised method of self-training, and to regression models after multivariate imputation by chain equation (MICE) of the unlabeled data. The models were developed for the most common toxicity outcomes in HNC patients, xerostomia and dysphagia, measured at six months after treatment, in a development cohort of 750 HNC patients. The models were externally validated in a validation cohort of 395 HNC patients. It was found that MICE and self-training did not have a gain in performance in terms of discrimination or calibration at external validation compared to current regression models. Therefore, the addition of unlabeled patient data by using the semi-supervised method of self-training or MICE would not be preferred in current NTCP modeling for HNC patients.

Introduction

Head and neck cancer (HNC) patients who undergo radiotherapy treatment often suffer from toxicities such as xerostomia (dry mouth) and dysphagia (difficulty swallowing) (Langendijk *et al.*, 2008). As these toxicities are long-lasting and impairing the quality of life in HNC patients, it is important that the toxicities are prevented by patient-specific clinical decision making. The use of Normal Tissue Complication Probability (NTCP) modeling can aid in these decisions, by describing the relationship between irradiation of normal tissue and the risk to develop toxicities (Kierkels *et al.*, 2014).

In NTCP modeling, often (logistic) regression models are used to predict toxicities in HNC patients. Recently, such models have been developed to make predictions for HNC patients in the decision to assign them to a new radiotherapy using protons (Langendijk *et al.*, 2013). This new therapy can - for certain patients - reduce the risk of radiation-induced toxicities, but due to practical constraints only a selective number of patients can be assigned to this therapy. NTCP modeling is therefore used to decide which patients have a large probability of developing severe toxicities and should therefore be prioritized in the selection for proton therapy. The most common toxicities, xerostomia and dysphagia, measured at six months after the start of radiotherapy treatment are modelled. Whether the expected reduction in complication risk is sufficiently lower for proton therapy is determined by using the modelled probabilities to compute the ΔNTCP value (see Langendijk *et al.*, 2013 and LIPPv2.2 for further details).

By using logistic regression in NTCP modeling, only patients from whom the toxicity outcomes have been recorded, the labelled data, are used to develop the model. However, as toxicities often arise long after the radiotherapy treatment has started (Langendijk *et al.*, 2008), a number of unlabeled observations is also available, in which the patient's background and treatment information has been documented, but not yet their toxicity outcomes. These unlabeled data are not used in logistic regression models, but the information that is present in these observations may possibly be of use to gain performance in NTCP modeling.

To include unlabeled data in NTCP modeling, semi-supervised learning techniques can be used. Semi-supervised learning is a type of machine learning that is conceptually between supervised and unsupervised learning (Van Engelen & Hoos, 2021). It is able to deal with both labelled and often large amounts of unlabeled data. However, also smaller amounts of unlabeled data can be used, as long as it conveys useful information for model development (Van Engelen & Hoos, 2021). Assumptions to be met for semi-supervised learning are the smoothness assumption (observations with the similar input should have the same labels), the low-density assumption (the decision boundary for the labels should pass through low-density area in the input space), and the manifold assumption (observations that are on the same low-dimensional manifold should have the same labels) (Chapelle *et al*., 2006; Van Engelen & Hoos, 2021).

Even though these techniques are able to incorporate both labelled and unlabeled data, they do not always necessarily result in better performance of the model compared to supervised learning. Moreover, it is difficult to determine a priori in which cases semi-supervised methods perform better than supervised methods. Depending on the specific dataset at hand, semi-supervised methods may improve or degrade in performance compared to their supervised equivalents (Chapelle *et al.*, 2006). In addition, it is difficult to determine which specific semi-supervised method works best for a given dataset and

research question (Van Engelen & Hoos, 2021). Therefore, in NTCP modeling to select HNC patients for proton therapy, it has yet to be evaluated whether the addition of unlabeled patient data by using semi-supervised methods would gain performance compared to the current (logistic) regression models, and if so, which semi-supervised method would be best.

One semi-supervised method that could be used in NTCP modeling is self-training, which is one of the oldest, widely known methods within semi-supervised learning (Triguero *et al*., 2015). The method was first proposed by Yarowsky (1995) and belongs to the so-called 'wrapper' methods within semi-supervised learning, which aim to work by 'pseudolabelling' the unlabeled data (Van Engelen & Hoos, 2021). In self-training, the model iteratively trains itself on the labelled data using a supervised classifier to assign pseudolabels to the unlabeled data. The supervised classifier or learner can be any method that classifies or predicts the labels and the classifier does not distinguish between the labeled and pseudolabeled data when developing the model (Van Engelen & Hoos, 2021). Each iteration, the model is re-trained using the labelled and pseudolabeled data (Van Engelen & Hoos, 2021). Beside the classifier or learner to include, a number of additional decisions to design a self-training model have to be made, such as the unlabeled data to label, the confidence threshold for the predictions to add pseudolabels, and the stopping criterion after which no more iterations take place (Triguero *et al.*, 2015). The confidence threshold determines the probabilities for the unlabeled outcomes that are certain enough to pseudolabel the outcomes; when the probability is larger than the threshold, the outcome is pseudolabeled as having the event, while outcomes with a probability smaller than 1 minus the threshold are labelled as not having the event. Outcomes that do not meet the threshold remain unlabeled in the respective iteration.

Self-training is computationally efficient and easily interpretable, making the method attractive to apply in clinical prediction settings as opposed to the more advanced semi-supervised methods that are available. Self-training has already been used previously within NTCP modeling for HNC patients treated at the Portuguese Institute of Oncology of Coimbra (Soares *et al.* 2016). The data contained 84 unlabeled observations of the 222 patients in total, and the use of self-training with a random forest classifier showed a gain in discrimination performance compared to supervised learning.

Here, it will be explored how the current supervised regression methods in NTCP modeling for proton therapy selection compare in performance to the semi-supervised method of self-training, when applied to a dataset of 750 observations of HNC patients treated at the University Medical Center Groningen. Of the 750 observations, 6 months of observations will be turned into unlabeled observations, reflecting the setting of data availability of the toxicity outcomes at six months in clinical practice. Six different methods will then be compared using the dataset containing labelled and unlabeled observations: (1) logistic regression, (2) ridge regression, (3) logistic regression after multiple imputation of the outcome with MICE, (4) ridge regression after multiple imputation of the outcome with MICE, (5) self-training with logistic regression as classifier, (6) self-training with ridge regression as classifier. As a difference in performance of the methods may be dependent on the size of the dataset, all models will additionally be applied to a range of datasets with decreasing numbers of labeled data.

Methods

Study design

The development cohort consisted of 750 observations from HNC patients treated at the University Medical Centre Groningen (between January 2007 and June 2016). The validation cohort consisted of 395 observations from patients treated at the University Medical Centre Groningen (between July 2016 and December 2017), Maastro Clinic (between May 2012 and June 2016), and Radiotherapeutic Institute Friesland (between May 2014 and December 2016).

In both cohorts, the observations included the presence of toxicities in HNC patients scored at different intervals during and after radiotherapy treatment. For each observation, the patient, tumor, and treatment characteristics and dose parameters of 28 organs were also included in the dataset. Further details on the content, collection, and inclusion requirements of the data have been previously described (Van den Bosch *et al.*, 2021).

Missing data

In the current dataset, the reasons of missing data can be divided into non-compliance, relapse, follow up too short, and death (Van den Bosch *et al.*, 2020). There are missing data present in baseline toxicity scores and toxicity outcomes as described in Table 1. In R, the function `mice()` within the "mice" package (Van Buuren & Groothuis-Oudshoorn, 2011) was used to impute the missing data in the development and validation cohort separately (and in the validation cohort also per center separately, from Van den Bosch *et al.*, 2021). The `mice()` function uses the multivariate imputation by chained equation (MICE) approach. With this procedure the missing data were filled in one time for the development cohort and ten times for the validation cohort to account for randomness. The multiple imputed datasets were used in further analyses and the results were pooled according to Rubin's rule (Rubin, 1987).

Outcomes

The focus of this study was on the most common toxicities present at six months after the end of treatment, which are xerostomia (dry mouth) and dysphagia (difficulty swallowing), as these are the outcomes currently used in the decision to assign patients to proton therapy (LIPPv2.2). Xerostomia is a patient-reported item ranging from grade 1 to 4 (1 = "not at all", 2 = "a little", 3 = "quite a bit", 4 = "very much") based on EORTC QLQ-H&N35 (question 41). Dysphagia is a physician-rated item ranging from grade 1 to 5 (1 = "symptomatic but normal diet", 2 = "only soft food", 3-5 = "liquid food or tube feeding") based on CTCAEv4.0 (see Van den Bosch *et al.*, 2021; and LIPPv2.2). A grade larger than 2 is considered as clinically relevant. The two toxicity outcomes were both dichotomized in two ways (grade ≥ 2 and grade ≥ 3), leading to a total of four outcomes that were modelled separately:

(a) xerostomia grade ≥ 2,

(b) xerostomia grade ≥ 3,

(c) dysphagia grade ≥ 2,

(d) dysphagia grade ≥ 3.

Patients who started their radiotherapy treatment shorter than six months ago are in practice unlabeled in the dataset, equal to approximately 40 patients. Therefore, to test the effect of unlabeled data in the practical setting of NTCP modeling, 40 random patient outcomes were made unlabeled.

Table 1 | Description of the development and validation cohorts.

	Development cohort (n=750)		Validation cohort (n=395)	
		Missing (%):		Missing (%):
Patient				
▪ Mean age (sd)	63.1 (10.2)	-	64 (9.4)	-
▪ Sex (%)				
Male	560 (75%)	-	290 (73%)	-
Female	190 (25%)	-	105 (27%)	-
▪ Tumor stage (%)				
Tis-T2	363 (48%)	-	194 (49%)	-
T3-T4	387 (52%)	-	201 (51%)	-
Mean dose to the (sd)				
▪ Submandibular glands	48.5 (22.9)	2 (0.3%)	44.7 (20.7)	-
▪ Parotid glands	51.7 (32.0)	-	51.7 (32.0)	-
▪ Pharyngeal constrictor muscle (PCM) superior	42.9 (24.1)	-	38.7 (23.)	-
▪ Pharyngeal constrictor muscle (PCM) medius	48.7 (20.3)	-	49.4 (20.1)	-
▪ Pharyngeal constrictor muscle (PCM) inferior	54.7 (13.0)	-	53.0 (14.1)	-
Toxicity at baseline (%)				
▪ Xerostomia grade ≥ 2	75 (11%)	85 (11%)	52 (18%)	99 (25%)
▪ Xerostomia grade ≥ 3	16 (2%)	85 (11%)	20 (7%)	99 (25%)
▪ Dysphagia grade ≥ 2	178 (24%)	14 (2%)	85 (22%)	-
▪ Dysphagia grade ≥ 3	62 (8%)	14 (2%)	21 (5%)	-
Toxicity at 6 months (%)				
▪ Xerostomia grade ≥ 2	260 (44%)	160 (21%)	93 (48%)	201 (51%)
▪ Xerostomia grade ≥ 3	78 (13%)	160 (21%)	30 (15%)	201 (51%)
▪ Dysphagia grade ≥ 2	183 (29%)	118 (16%)	61 (19%)	71 (18%)
▪ Dysphagia grade ≥ 3	94 (15%)	118 (16%)	21 (6%)	71 (18%)
Primary tumor location				
▪ Pharynx	372 (50%)	-	205 (52%)	-
▪ Larynx	334 (45%)	-	168 (43%)	-

Predictors

In Table 2 the predictors are listed for the xerostomia (grade ≥ 2 and grade ≥ 3) and dysphagia (grade ≥ 2 and grade ≥ 3) models respectively. These are a preselected subset of the variables that were present in or derived from the dataset and that were determined clinically relevant based on expertise knowledge and previous research. Dose variables were measured in Gray (Gy) and treated as continuous variables, while the presence of baseline toxicity and tumor site were treated as binary variables. The doses to both the submandibular glands were added up and combined into one predictor. The doses to the contralateral and ipsilateral parotid gland were each root transformed and then added up and combined into one predictor. These new definitions and transformations were in line with the latest version of the LIPP (v2.2). There were no other transformations or interactions of the predictors in the models. These same sets of predictors were used in all the different modeling methods described below.

Table 2 | The preselected set of predictors from the dataset that were used in the models to predict the presence of grade ≥ 2 and grade ≥ 3 xerostomia and dysphagia at 6 months after the end of radiotherapy in HNC patients.

Predictors for xerostomia (grade ≥ 2 or 3)	**Predictors for dysphagia (grade ≥ 2 or 3)**
Mean dose (Gy) to the (continuous) ▪ Submandibular glands ▪ Ipsilateral parotid gland (sqrt) + contralateral parotid gland (sqrt) Xerostomia at baseline (binary) ▪ Grade ≥ 2 ▪ Grade ≥ 3	Mean dose (Gy) to the (continuous) ▪ Pharyngeal constrictor muscle (PCM) superior ▪ Pharyngeal constrictor muscle (PCM) medius ▪ Pharyngeal constrictor muscle (PCM) inferior Dysphagia at baseline (binary) ▪ Grade ≥ 2 ▪ Grade ≥3 Primary tumor location (binary) ▪ Pharynx ▪ Larynx

Models

Six different models were created in R to compare the performance of supervised and semi-supervised methods and the inclusion of unlabeled data: (1) logistic regression, (2) ridge regression, (3) logistic regression after multiple imputation of the outcome with MICE, (4) ridge regression after multiple imputation of the outcome with MICE, (5) self-training with logistic regression as classifier, (6) self-training with ridge regression as classifier. These six models were developed separately for all four toxicity outcomes (xerostomia grade ≥ 2, xerostomia grade ≥3, dysphagia grade ≥ 2, and dysphagia grade ≥3), thus leading to a total of 24 models.

<u>Logistic and ridge regression models</u>

The multivariable logistic regression models and ridge regression models were used as supervised baselines. These models could only use the part of the development cohort that is labelled (i.e. the development cohort minus the observations of six months that were made unlabeled).

Logistic and ridge regression models with MICE

MICE was additionally used to impute the missing data in the development cohort that were missing because of un-labelling 40 of the patients' outcomes. After imputation, logistic and ridge regression models were created on the imputed dataset, thereby using a fully labelled dataset. In this way, the methods are still regarded as supervised, but do include the unlabeled data after imputation.

Logistic and ridge regression models with self-training

The semi-supervised method of self-training was used to examine whether the inclusion of unlabeled data by semi-supervised learning would improve NTCP modeling. By using self-training, the labelled data were used to train the model aiming to predict the labels for the unlabeled data. A confidence threshold determines the most confident predictions for the unlabeled data. With high thresholds, the self-training method may not select many unlabeled data to be added to the data set, resulting in no improvement compared to the supervised baseline. Low thresholds may add wrongly classified unlabeled data to the dataset, resulting in a decrease in performance compared to the supervised baseline. The confidence threshold was therefore set at an intermediate value of 0.8 (following Soares *et al.* (2016) in self-training models for xerostomia prediction). The process of model development and using the predictions to add labels was iterated until it reached a stopping criterion, which was determined to be when no unlabeled data were left or after a maximum of 50 iterations.

Model performance

The models were externally validated using the validation cohort of 395 labelled observations. Both the discrimination (the ability to differentiate between the outcomes) and the calibration (the consistency between predicted and actual probability of the outcome) of the models was evaluated.

For the discrimination, the area under the ROC curve (AUC) was used, which is the surface area under the curve that plots 1 - specificity against the sensitivity. It indicates how well the model can distinguish between individuals with and without the outcome based on their predicted risk. The AUC is a number between 0 and 1 with a higher AUC value indicating better discrimination of the model. For the calibration, the following measures were determined:

(1) 'mean calibration' (or 'calibration-in-the-large') which is the average predicted probability compared with the overall outcome rate,

(2) 'weak calibration' by calculating the calibration slope, and

(3) 'moderate calibration' which visually shows how the estimated probabilities correspond to observed proportions with a calibration curve (closeness to the diagonal).

At external validation, these measures indicate whether the model has good generalizability (i.e. the ability to accurately predict outcomes for HNC patients from different but related populations). The `val.prob.ci.2()` function (Van Calster *et al.*, 2016) was used to derive the above calibration performance measures.

Ratio labelled vs. unlabeled data

In addition, it was evaluated whether differences in performance within or between the different models are dependent on the number of labelled observations in the dataset. Therefore, in a separate part of the analysis the number of observations was decreased stepwise. It was determined to keep the introduction of unlabeled data fixed at the average number of new patients each six months, which equals approximately 40 patients in this dataset, as this is most similar to the data availability in NTCP modeling for HNC patients in practice. The amount of labelled data was then decreased stepwise by this number, by each time additionally removing 40 random observations from the dataset (Table 3).

After each decrease in number of observations in the dataset, all six models were again independently developed for the four outcomes. For the regular logistic and ridge regression models (model 1 and 2), this meant that only the labelled observations that were left, were included in the development of the model. The other models did include the 40 unlabeled observations, either by imputation after MICE (model 3 and 4) or by pseudolabelling with self-training (model 5 and 6).

Table 3 | Amounts of labelled and unlabeled observations to compare the performance of the methods at different amounts of labelled data, including the corresponding absolute number of events and events per variable (EPV) for each of the four modelled outcomes respectively.

	Xerostomia grade ≥ 2		Xerostomia grade ≥ 3		Dysphagia grade ≥ 2		Dysphagia grade ≥ 3	
	Events	EPV	Events	EPV	Events	EPV	Events	EPV
710 labelled + 40 unlabeled	323	81	94	24	236	30	127	16
670 labelled + 40 unlabeled	303	76	90	22	220	28	119	15
630 labelled + 40 unlabeled	288	72	88	22	205	26	110	14
590 labelled + 40 unlabeled	271	68	85	21	193	24	103	13
550 labelled + 40 unlabeled	254	64	75	19	179	22	95	12
510 labelled + 40 unlabeled	236	59	70	18	166	21	87	11
470 labelled + 40 unlabeled	221	55	64	16	149	19	77	10
430 labelled + 40 unlabeled	202	50	60	15	134	17	67	8
390 labelled + 40 unlabeled	182	46	53	13	117	15	57	7
350 labelled + 40 unlabeled	165	41	48	12	104	13	51	6
310 labelled + 40 unlabeled	150	38	45	11	93	12	44	6
270 labelled + 40 unlabeled	129	32	40	10	81	10	37	5
230 labelled + 40 unlabeled	110	28	35	9	70	9	31	4
190 labelled + 40 unlabeled	89	22	-	-	-	-	-	-
150 labelled + 40 unlabeled	66	16	-	-	-	-	-	-
110 labelled + 40 unlabeled	48	12	-	-	-	-	-	-
70 labelled + 40 unlabeled	32	8	-	-	-	-	-	-

Results

The six models were externally validated in the validation cohort (n=395). For each of the four outcomes, the external validation results are described below. The calibration curves and the regression coefficients of all models for the four outcomes are presented in Appendix Figures A1-A4 and Tables A1-A4, respectively. The number of iterations and pseudolabels added by the two self-training models are presented in Appendix Tables A5-A8.

Model performances for xerostomia grade ≥ 2

The discrimination and calibration at external validation of the xerostomia grade ≥ 2 models with 710 labelled and 40 unlabeled observations are presented in Table 4. The models appear to show similar performance in predicting xerostomia ≥ 2. The three ridge regression models show values for calibration-in-the-large that are only slightly closer to the ideal value of zero, and values for the calibration slopes that are slightly closer to the ideal value of 1, as opposed to the three logistic regression models. The similarity of the six models is also apparent in the calibration curves (Appendix Figure A1), in which the shape and closeness to the diagonal of the curves are similar across all models.

Table 4 | External validation of the xerostomia grade ≥ 2 models with 710 labelled and 40 unlabeled observations.

Model	AUC (standard error)	Calibration-in-the-large (standard error)	Calibration slope (standard error)
Logistic regression	0.68 (0.027)	0.20 (0.109)	0.79 (0.133)
Ridge regression	0.68 (0.027)	0.18 (0.108)	0.87 (0.145)
MICE (logistic regression)	0.67 (0.027)	0.18 (0.110)	0.74 (0.127)
MICE (ridge regression)	0.68 (0.027)	0.16 (0.109)	0.83 (0.139)
Self-training (logistic regression)	0.68 (0.027)	0.20 (0.109)	0.78 (0.131)
Self-training (ridge regression)	0.68 (0.027)	0.18 (0.108)	0.86 (0.144)

Model performances for xerostomia grade ≥ 2 with decreasing data

To test whether the performance of the six models is dependent on the amount of data, the number of labeled observations in the dataset was decreased by steps of 40 observations. In Figure 1a, the AUCs of the different xerostomia grade ≥ 2 models appear to be almost equal and remain at a constant level when less data is used. However, when the models include less then around 150 labelled observations, the AUCs of the three logistic models show a steep decrease, while the three ridge models show a slight increase. None of the six models has a clearly higher AUC overall, but the three models involving ridge regression have slightly higher AUCs across smaller numbers of labelled observations.

With regard to the calibration-in-the-large of the xerostomia grade ≥ 2 models, depicted in Figure 1b, there is again no clear difference between the six models until the amount of labelled data is decreased towards 200 observations and the three logistic regression models tend to perform worse. For the calibration slope, Figure 1c shows that for large numbers of observations, the six models tend to perform similar. When the amount of observations decreases, the three models involving logistic regression have a better slope value closer to 1.

The general absence of a clear difference between the regression models with or without self-training, can be due to the small number of pseudolabels (only 4 to 5) that were added by the self-training models for the xerostomia grade ≥ 2 outcome (Appendix Table A5).

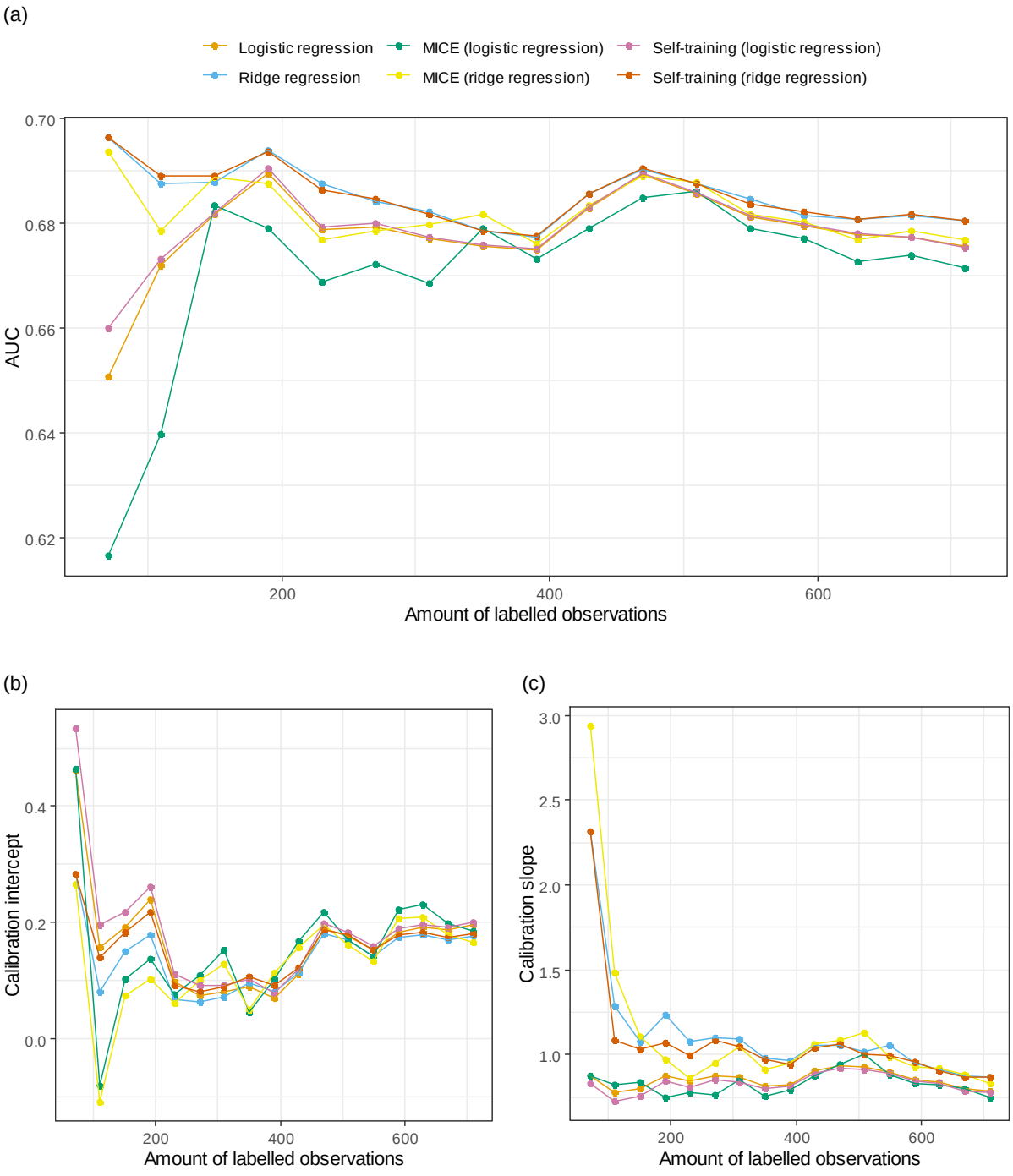


Figure 1 | External validation of the models for xerostomia grade ≥ 2 models for different amounts of labeled data. The amount of labelled data is fixed at 40 observations. (a) The AUCs, (b) the calibration intercepts, and (c) the calibration slopes.

Model performances for xerostomia grade ≥ 3

For the xerostomia grade ≥ 3 models, the external validation of the models with 710 labelled and 40 unlabeled observations are shown in Table 5. Again the six models seem similar in performance, with the ridge regression only having slightly higher calibration-in-the-large and calibration slope values.

Table 5 | External validation of the xerostomia grade ≥ 3 models with 710 labelled and 40 unlabeled observations.

Model	AUC (standard error)	Calibration-in-the-large (standard error)	Calibration slope (standard error)
Logistic regression	0.69 (0.036)	0.23 (0.142)	0.80 (0.184)
Ridge regression	0.69 (0.035)	0.26 (0.141)	0.90 (0.201)
MICE (logistic regression)	0.69 (0.036)	0.25 (0.142)	0.81 (0.185)
MICE (ridge regression)	0.69 (0.035)	0.27 (0.141)	0.90 (0.202)
Self-training (logistic regression)	0.69 (0.036)	0.27 (0.142)	0.79 (0.180)
Self-training (ridge regression)	0.70 (0.035)	0.31 (0.140)	0.95 (0.213)

Model performances for xerostomia grade ≥ 3 with decreasing data

Also, similar to the xerostomia grade ≥ 2 models, the xerostomia grade ≥ 3 models involving ridge regression show only slightly higher AUCs compared to logistic regression across the different amounts of labelled data (Figure 2a). But when the amount of observations is decreased towards 350 labelled observations, the higher AUC of the ridge regression models becomes more apparent.

With regard to the calibration-in-the-large across different amounts of labelled observations, a more stable pattern is visible for the six different models (Figure 2b). The two MICE models and the regular logistic regression models have calibration-in-the-large values that are closest to zero across all amounts of observations. For the calibration slope, all models show a similar pattern until around 450 observations, when the three ridge regression models have slopes increasingly further away from 1 (Figure 2c).

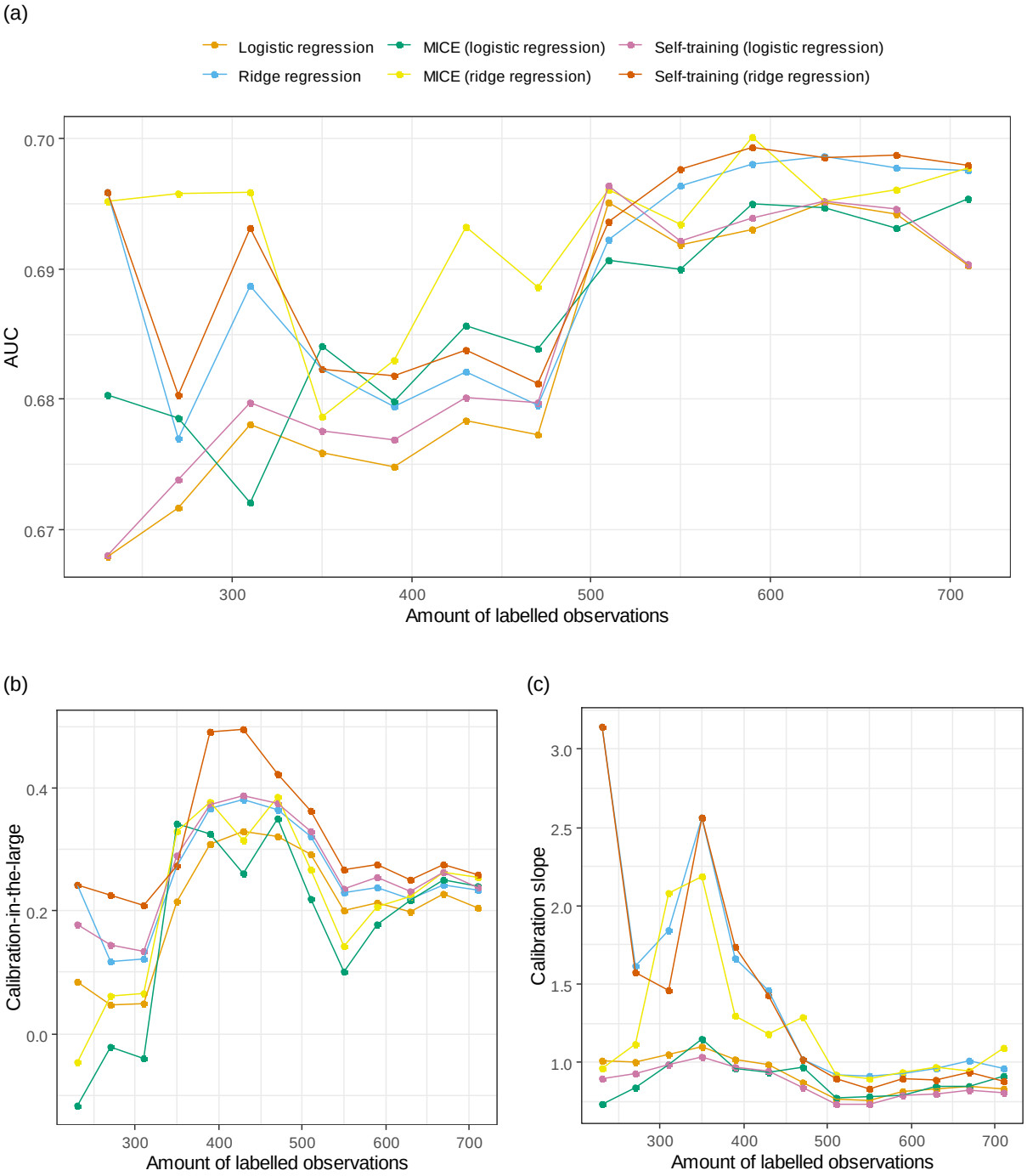


Figure 2 | External validation of the models for xerostomia grade ≥ 3 models for different amounts of labeled data. The amount of labelled data is fixed at 40 observations. (a) The AUCs, (b) the calibration intercepts, and (c) the calibration slopes.

Model performances for dysphagia grade ≥ 2

For the dysphagia grade ≥ 2 models, the external validation for the six models with 710 and 40 labelled observations are shown in Table 6. Though the AUC does not differ between the models, the calibration-in-the-large is closer to zero, and the calibration slope is closer to 1, for the three models involving ridge regression. This is also visible from the calibration curves of the dysphagia grade ≥ 2 models (Appendix Figure A3). All calibration curves of the ridge regression models are closer to the diagonal compared to the logistic regression models.

Table 6 | External validation of the dysphagia grade ≥ 2 models with 710 labelled and 40 unlabeled observations.

Model	AUC (standard error)	Calibration-in-the-large (standard error)	Calibration slope (standard error)
Logistic regression	0.74 (0.026)	0.61 (0.134)	0.60 (0.086)
Ridge regression	0.74 (0.026)	0.24 (0.128)	0.75 (0.108)
MICE (logistic regression)	0.74 (0.026)	0.55 (0.134)	0.60 (0.086)
MICE (ridge regression)	0.74 (0.026)	0.21 (0.128)	0.74 (0.107)
Self-training (logistic regression)	0.74 (0.026)	0.64 (0.135)	0.59 (0.084)
Self-training (ridge regression)	0.74 (0.026)	0.25 (0.128)	0.74 (0.107)

Model performances for dysphagia grade ≥ 2 with decreasing data

When decreasing the number of observations for the dysphagia grade ≥ 2 models in Figure 3a, it becomes visible that all models have similar AUC values and decrease almost in the same way. However, the AUC is often lower for the three ridge regression models.

With regard to the calibration-in-the-large, presented in Figure 3b, the logistic and ridge models for dysphagia grade ≥ 2 are clearly different. The three models involving ridge regression have calibration-in-the-large values closest to zero across all numbers of observations, and do not overlap with the higher intercept values of the models involving logistic regression. The same is true for the calibration slope presented in Figure 3c, where the three models involving ridge regression have calibration slope values closest to 1.

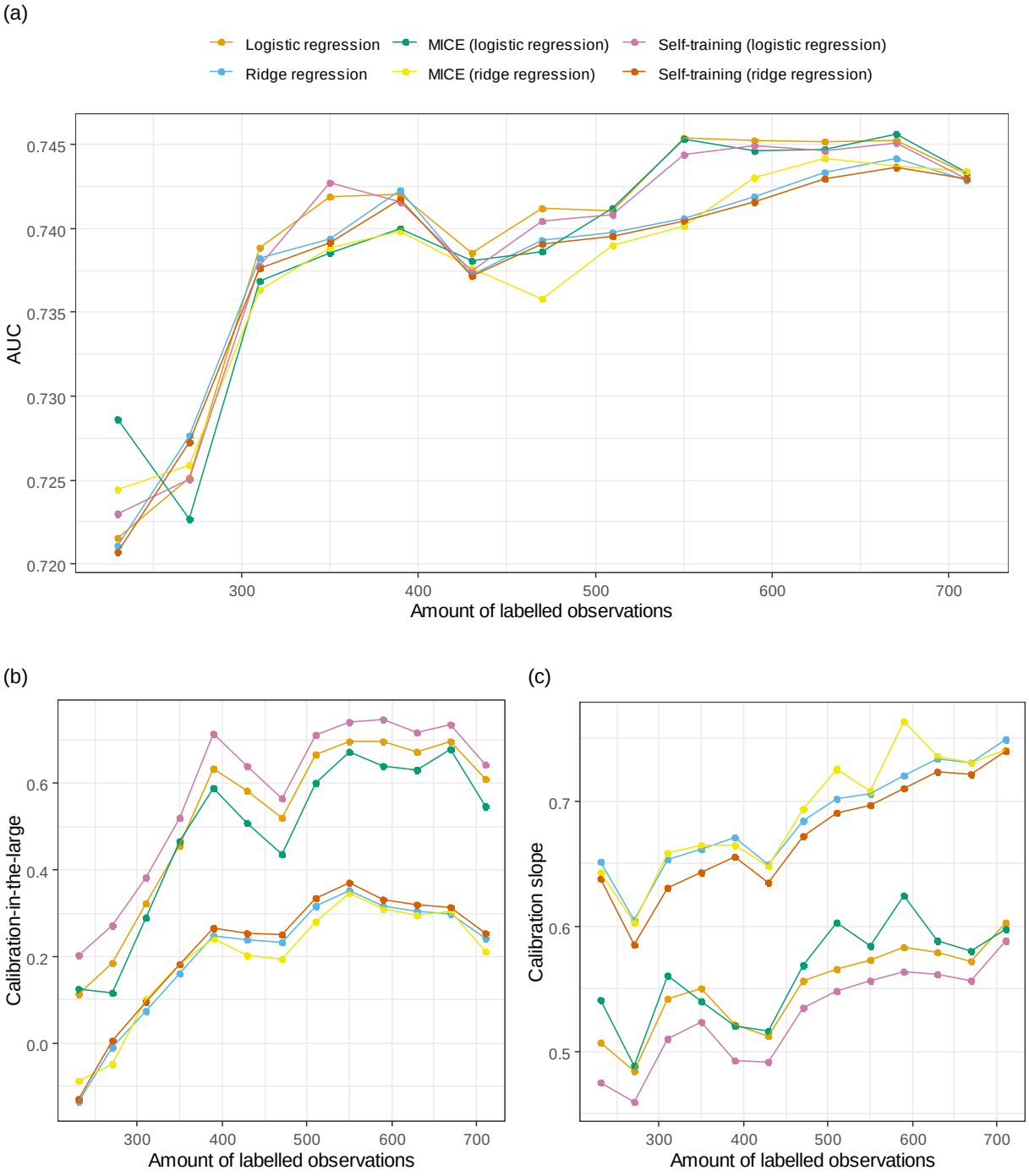


Figure 3 | External validation of the models for dysphagia grade ≥ 2 models for different amounts of labeled data. The amount of labelled data is fixed at 40 observations. (a) The AUCs, (b) the calibration intercepts, and (c) the calibration slopes.

Model performances for dysphagia grade ≥ 3

For the dysphagia grade ≥ 3 models, the external validation for the models using 710 labelled and 40 unlabeled observations is shown in Table 7. The AUC values for the three models involving ridge regression is only slightly lower than for the models involving logistic regression. The calibration-in-the-large however, is closer to zero for the models using ridge regression, and also the calibration slope is closer to 1 compared to the models with logistic regression. The slopes in general, however, appear to be rather low (between 0.44 and 0.55) for all six models. This is also visible from the calibration curves, in which all models show slight underprediction for low observed proportions and overprediction for high observed proportions (Appendix Figure A4).

Table 7 | External validation of the dysphagia grade ≥ 3 models with 710 labelled and 40 unlabeled observations.

Model	AUC (standard error)	Calibration-in-the-large (standard error)	Calibration slope (standard error)
Logistic regression	0.74 (0.038)	0.49 (0.181)	0.46 (0.097)
Ridge regression	0.72 (0.040)	0.07 (0.174)	0.55 (0.119)
MICE (logistic regression)	0.73 (0.037)	0.55 (0.181)	0.44 (0.097)
MICE (ridge regression)	0.72 (0.039)	0.12 (0.172)	0.55 (0.123)
Self-training (logistic regression)	0.74 (0.038)	0.57 (0.182)	0.45 (0.095)
Self-training (ridge regression)	0.72 (0.040)	0.10 (0.174)	0.54 (0.118)

Model performances for dysphagia grade ≥ 3 with decreasing data

When decreasing the amount of observations, a clear distinction in AUC is visible for the six dysphagia ≥ 3 models (Figure 4a). The three models involving ridge regression have a constant lower AUC value across all amount of observations. Furthermore, the AUCs of the models using ridge regression with or without self-training remain almost identical when decreasing the amount of observations. For logistic regression with or self-training, however, the self-training performed better across all numbers of observations.

With regard to the calibration-in-the-large, the models with ridge regression have intercept values closer to zero at larger amounts of observations, but when the amount of observation decreases, the intercept value of the logistic regression models is closer to zero (Figure 4b). For the calibration slope, the three models involving ridge regression seem to have a slope constantly closer to 1 compared to the logistic models (Figure 4c). The slopes of the self-training models with logistic or ridge regression perform slightly worse compared to their supervised baseline of regular logistic and ridge regression respectively.

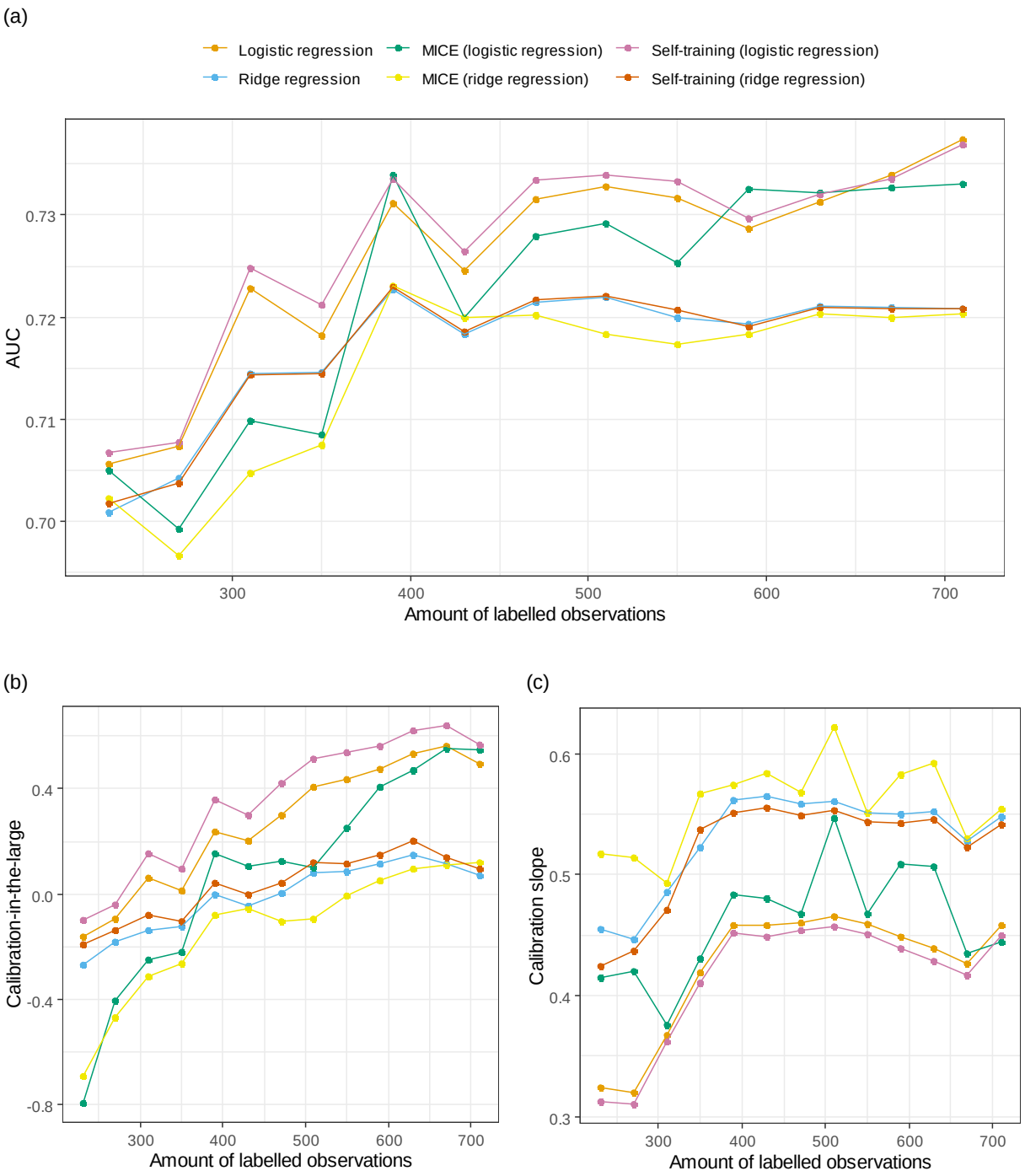


Figure 4 | External validation of the models for dysphagia grade ≥ 3 models for different amounts of labeled data. The amount of labelled data is fixed at 40 observations. (a) The AUCs, (b) the calibration intercepts, and (c) the calibration slopes.

Self-training: confidence threshold

As no substantial differences were found between the models with or without self-training, it was determined to test the effect of the confidence threshold on the performance of the self-training method in this dataset. Therefore, self-training with logistic regression was repeated in an additional analysis for the xerostomia grade ≥ 2 outcome using confidence thresholds of 0.5, 0.6, 0.7, 0.8, 0.9, and 0.95, and then compared to regular logistic regression and logistic regression after MICE. It was found that with lower confidence thresholds, more pseudolabels were added during the self-training process, and for confidence thresholds above 0.8 only few of the 40 unlabeled data were pseudolabeled to almost none when the confidence threshold was set at 0.9 or larger (Appendix Table A9). However, with regard to the performance of the self-training at external validation, none of the tested thresholds used in self-training was related to a better performance compared to the regular logistic regression method (Appendix Figure A5).

Self-training: ratio of labelled/unlabeled data

Since the number of 40 unlabeled observations that was introduced may be relatively low, it was determined to examine the effect of larger proportions of unlabeled observations on the performance of the six different methods. In another additional analysis, the number of labelled observations was decreased by steps of 40, while the total number of observations was kept at 750. It was found that when the proportion of unlabeled data in the dataset was larger, the performance of the two self-training methods was slightly better in terms of discrimination (AUC), but worse in terms of calibration (calibration-in-the-large and calibration slope) compared to the other four methods when externally validated (Appendix Figure A6). The calibration performance of the two self-training methods started to decrease below 600 observations, the point at which the methods started to add incorrectly classified pseudolabels (Appendix Table A10).

Discussion

The aim of this study was to compare the currently used regression methods in NTCP modeling to the semi-supervised method of self-training and to regression after MICE, in order to examine the possible gain in performance when additionally using unlabeled data in model development. The results show that none of the six tested models had the best discrimination (AUC) and calibration (calibration-in-the-large and calibration slope) across all separately modelled outcomes (xerostomia ≥ grade 2, xerostomia ≥ grade 3, dysphagia grade ≥ 2, dysphagia grade ≥ 3) when externally validated. Across different amounts of labelled data, self-training with logistic regression or ridge regression tended to perform similar or even slightly worse than the regular logistic or ridge regression models. However, overall the three models based on ridge regression showed slightly better discrimination and/or calibration compared to the three models with logistic regression, and for some the dysphagia outcomes this trend was even more apparent.

Whether self-training would gain in model performance compared to supervised learning, depends on the dataset and specific research question at hand (Van Engelen & Hoos, 2021). In the

current NTCP modeling, relatively few of the toxicity outcomes are unlabeled as only for most recent patients the toxicity outcomes have not yet documented, which in this case is equal to about 40 patients within a dataset that has already over 700 labelled observations. Overall, across the different sizes of labelled data all models seemed to gain or lose in performance in the same way, and the relative performance among the models did not clearly change across the sizes of labelled data. The reason that self-training did not show better calibration or discrimination in this dataset, may be the relatively small amount of unlabeled data. In the study by Soares *et al.* (2016), self-training did show a gain in discriminative performance when the dataset had a larger proportion of unlabeled data (87 out of 222 observations), and also Chi *et al.* (2019) found better discrimination and calibration performance when the amount of unlabeled data increased in a dataset of over 100.000 observations of the survival of colorectal cancer patients. In the additional analysis performed in the current study, it also showed that when the proportion of unlabeled data was larger, the self-training methods performed slightly better in terms of discrimination, but not calibration. This may be due to the addition of incorrect pseudolabels that does not affect the model's ability to discriminate between having the event or not, but may affect the exact predicted probabilities. Nevertheless, larger proportions of data are not representative for the practical setting of NTCP modeling of HNC patients.

More important than the number of unlabeled data included in the dataset, may be the information that is conveyed by the unlabeled observations (Van Engelen & Hoos, 2021), and the distribution of the examples in the classification problem (Chapelle *et al.*, 2006). When more overlap between the two classes of the outcome is present, the self-training may have more difficulty providing the correct pseudolabels and may therefore be impaired in its performance. This may have caused the similarities and degradations in performance of the self-training models compared to the logistic regression models in the current dataset.

An increase in the confidence threshold could prevent the addition of wrong pseudolabels, as the choice of the confidence threshold can significantly influence the performance of self-training (Van Engelen & Hoos, 2021). For the xerostomia grade ≥ 2 outcome in the current dataset, it was found that the confidence threshold determined how many (incorrect) pseudolabels were added, but not how well the model performed. It could be further examined whether changes in the confidence threshold of the self-training method would improve the performance of the self-training for different outcomes, and additionally, if and how the optimal confidence threshold varies for different (number of events in the) outcomes.

Even though self-training did not show a gain in performance for this dataset, the three models that used ridge regression did sometimes show a better performance compared to the three logistic models. Especially for the dysphagia models this difference was more apparent. For the dysphagia outcomes, the number of predictors was larger (eight for dysphagia compared to four for xerostomia) and the predictors themselves were different. These may have played a role in the differences in discrimination and calibration between ridge and logistic regression. Ridge regression is aimed to prevent overfitting with larger number of predictors and deals with multicollinearity, which is likely more applicable for the dysphagia predictors, but this was not further examined. Ridge regression has been used in xerostomia prediction models before (e.g. Han *et al.*, 2019), but the interpretability and clinical

usefulness may be less beneficial compared to logistic regression models that use other ways to deal with multicollinearity (Van den Bosch *et al*., 2020)

In conclusion, the addition of unlabeled data in NTCP modeling by semi-supervised self-training did not lead to a better performance in terms of discrimination and calibration. The exact performance of self-training was similar or worse compared to logistic regression, depending on the specific outcome that was modelled. Furthermore, it was shown that for the xerostomia grade ≥ 2 outcome the confidence threshold of the self-training was not related to the performance of the self-training, but an increase in the ratio of unlabeled data did lead to a slightly better discriminative performance. Beside the performance of the model, other motivations to choose a model are the easy interpretability and applicability in practice. Of the six models tested here, the regression methods are most easily interpretable and applicable, and since self-training or MICE did not show a clear gain in performance, regression may be favored in practice.

References

Chapelle, O., Schölkopf, B., & Zien, A. (2006). Semi-supervised learning (1st ed.). Cambridge: The MIT Press

Chi, S., Li, X., Tian, Y., Li, J., Kong, X., Ding, K., Weng, C., & Li, J. (2019). Semi-supervised learning to improve generalizability of risk prediction models. *Journal of Biomedical Informatics*, *92*, 103117.

Han, P., Lakshminarayanan, P., Jiang, W., Shpitser, I., Hui, X., Lee, S. H., Cheng, Z., Guo, Y., Taylor, R. H., Siddiqui, S. A., Bowers, M., Sheikh, K., Kiess, A., Page, B. R., Lee, J., Quon, H. & McNutt, T. R. (2019). Dose/Volume histogram patterns in Salivary Gland subvolumes influence xerostomia injury and recovery. *Scientific reports*, *9*(1), 1-9.

Kierkels, R. G., Korevaar, E. W., Steenbakkers, R. J., Janssen, T., van't Veld, A. A., Langendijk, J. A., Schilstra, C, & van der Schaaf, A. (2014). Direct use of multivariable normal tissue complication probability models in treatment plan optimisation for individualised head and neck cancer radiotherapy produces clinically acceptable treatment plans. *Radiotherapy and Oncology*, *112*(3), 430-436.

Landelijk Platform Protonentherapie (LPPT) & Landelijk Platform Radiotherapy Hoofd-halstumoren (LPRHHT) (2019). Landelijk Indicatie Protocol Protonentherapie, versie 2.2 (LIPPv2.2). https://nvro.nl/publicaties/rapporten

Langendijk, J. A., Doornaert, P., Verdonck-de Leeuw, I. M., Leemans, C. R., Aaronson, N. K., & Slotman, B. J. (2008). Impact of late treatment-related toxicity on quality of life among patients with head and neck cancer treated with radiotherapy. *Journal of clinical oncology*, *26*(22), 3770-3776.

Langendijk, J. A., Lambin, P., De Ruysscher, D., Widder, J., Bos, M., & Verheij, M. (2013). Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach. *Radiotherapy and Oncology*, *107*(3), 267-273.

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons

Soares, I., Dias, J., Rocha, H., Khouri, L., Carmo Lopes, M. D., & Ferreira, B. (2016). Semi-supervised self-training approaches in small and unbalanced datasets: Application to xerostomia radiation side-effect. In *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016* (pp. 828-833). Springer, Cham.

Triguero, I., García, S., & Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. Knowledge and Information Systems, 42(2), 245–284.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. https://www.jstatsoft.org/v45/i03/

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M.J., Steyerberg E.W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 74, 167-176.

Van den Bosch, L., Schuit, E., van der Laan, H. P., Reitsma, J. B., Moons, K. G. M., Steenbakkers, R. J. H. M., Langendijk, J. A., & van der Schaaf, A. (2020). Key challenges in normal tissue complication probability model development and validation: towards a comprehensive strategy. *Radiotherapy and Oncology*, *148*, 151-156.

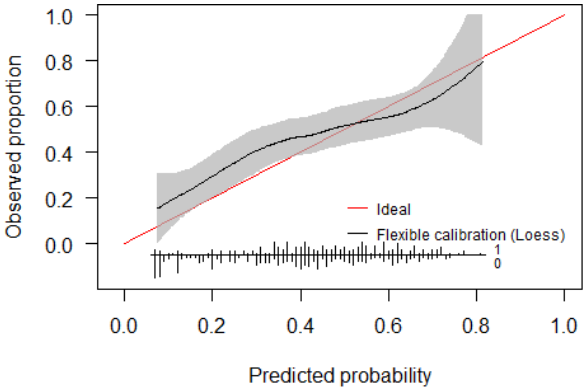
Van den Bosch, L., van der Schaaf, A., van der Laan, H. P., Hoebers, F. J. P., Wijers, O. B., van den Hoek, J. G. M., Moons, K. G. M., Reitsma, J. B., Steenbakkers, R. J. H. M, Schuit, E., & Langendijk, J. A. (2021). Comprehensive toxicity risk profiling in radiation therapy for head and neck cancer: A new concept for individually optimised treatment. *Radiotherapy and Oncology*, *157*, 147-154.

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning, 109*(2), 373-440.

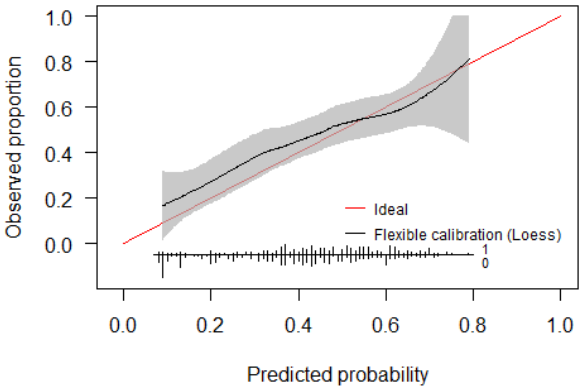
Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd annual meeting of the association for computational linguistics, association for computational linguistics (pp. 189–196).

Appendix

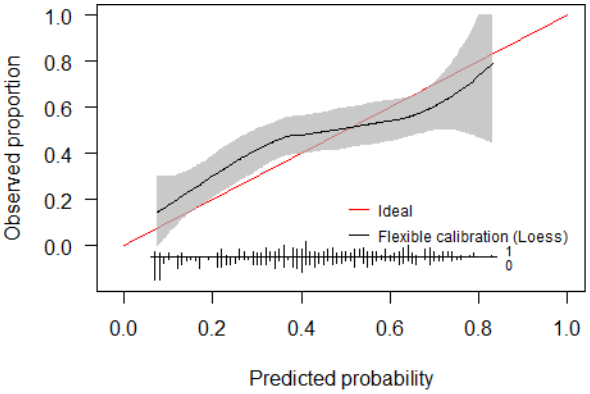
(a) Logistic regression



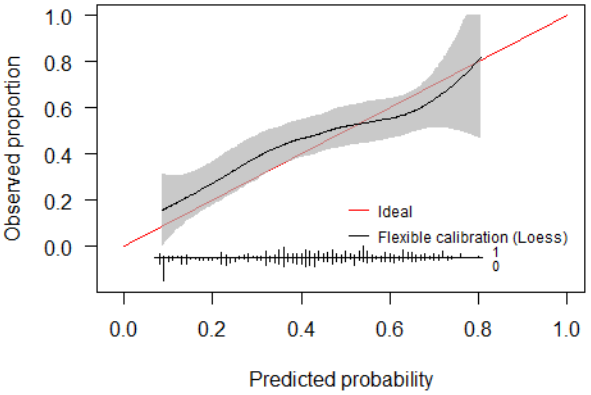
(b) Ridge regression



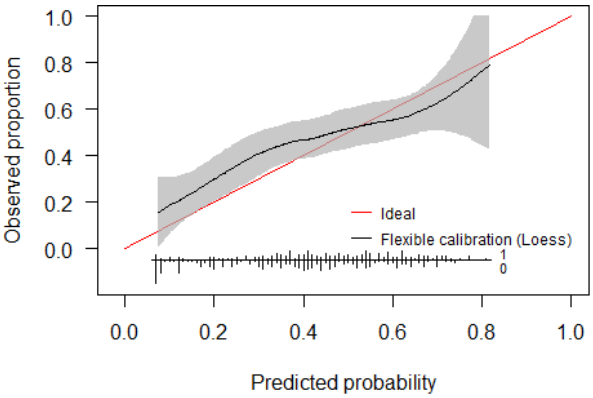
(c) Logistic regression after MICE



(d) Ridge regression after MICE



(e) Self-training with logistic regression as classifier



(f) Self-training with ridge regression as classifier

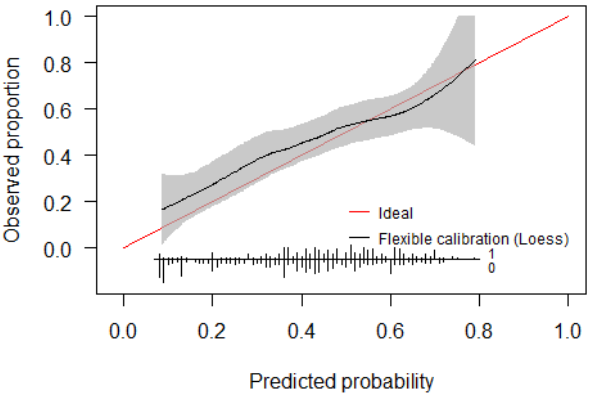
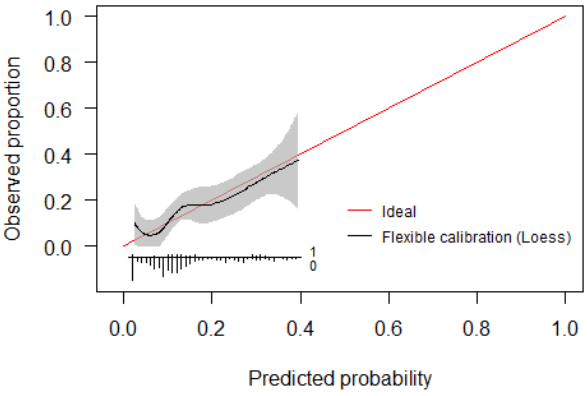
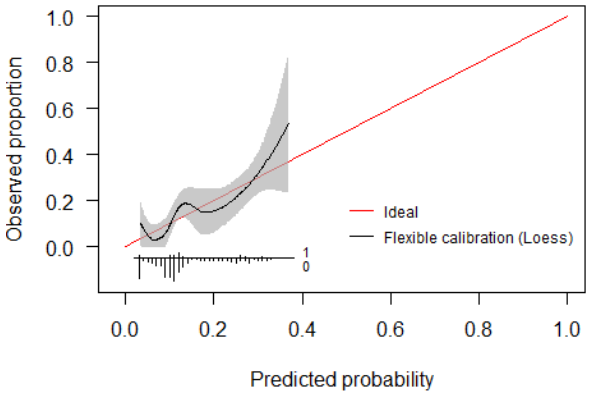


Figure A1 | Calibration curves of the xerostomia grade ≥ 2 models. (a) Logistic regression, (b) Ridge regression, (c) Logistic regression after multiple imputation of the outcome with MICE, (d) Ridge regression after multiple imputation of the outcome with MICE, (e) Self-training with logistic regression as classifier, (f) Self-training with ridge regression as classifier.

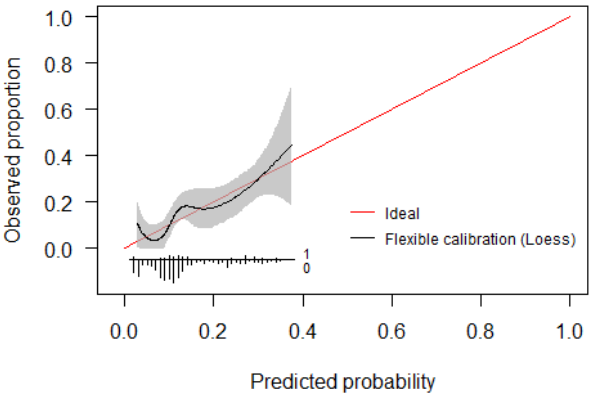
(a) Logistic regression



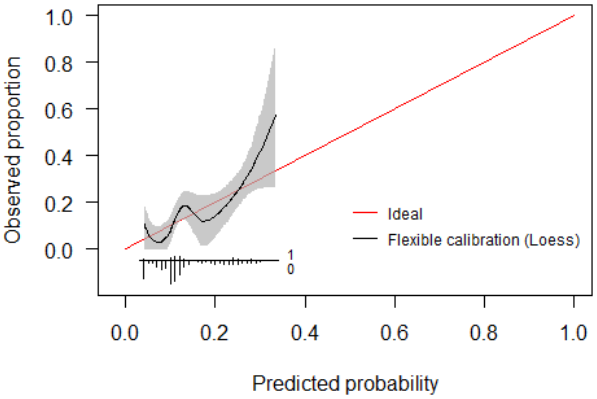
(b) Ridge regression



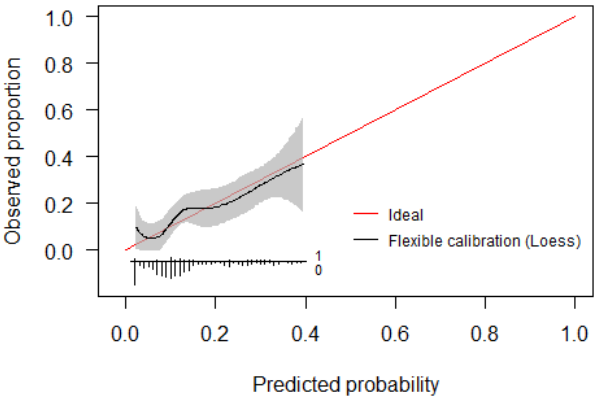
(c) Logistic regression after MICE



(d) Ridge regression after MICE



(e) Self-training with logistic regression as classifier



(f) Self-training with ridge regression as classifier

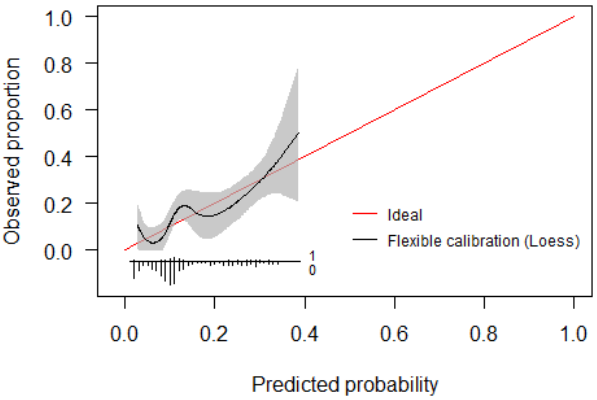
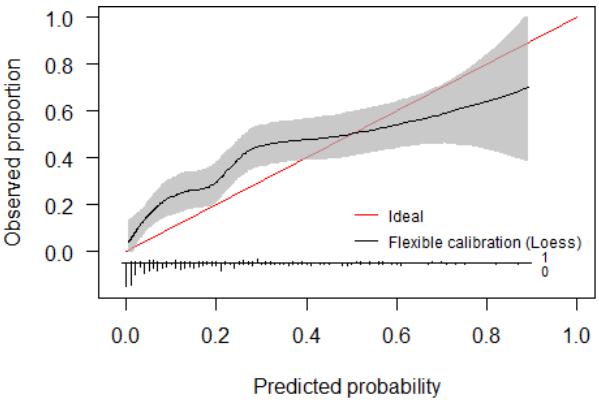
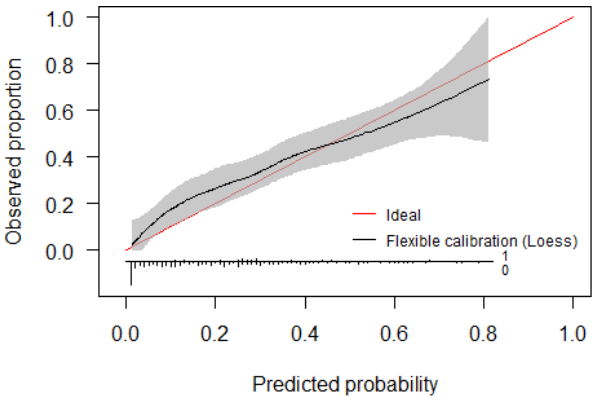


Figure A2 | Calibration curves of the xerostomia grade ≥ 3 models. (a) logistic regression, (b) ridge regression, (c) logistic regression after multiple imputation of the outcome with MICE, (d) ridge regression after multiple imputation of the outcome with MICE, (e) self-training with logistic regression as classifier, (f) self-training with ridge regression as classifier.

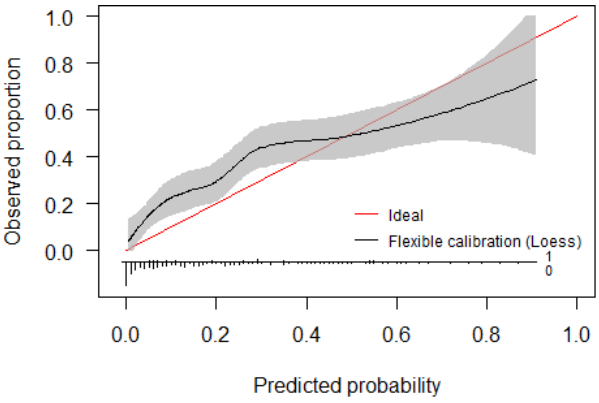
(a) Logistic regression



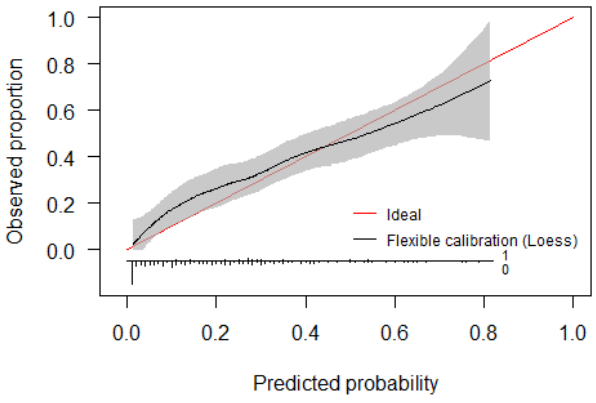
(b) Ridge regression



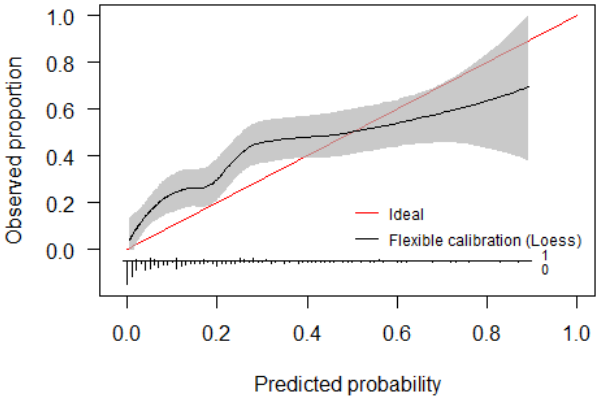
(c) Logistic regression after MICE



(d) Ridge regression after MICE



(e) Self-training with logistic regression as classifier



(f) Self-training with ridge regression as classifier

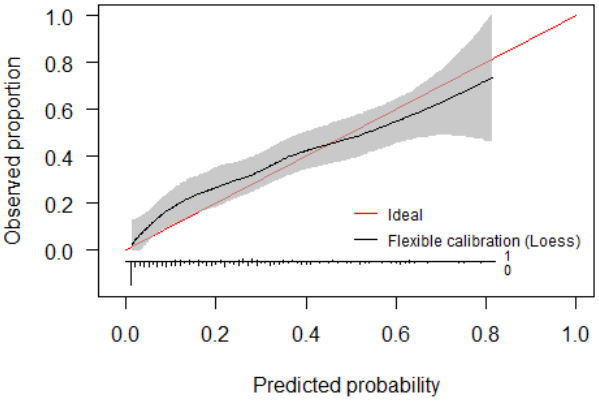
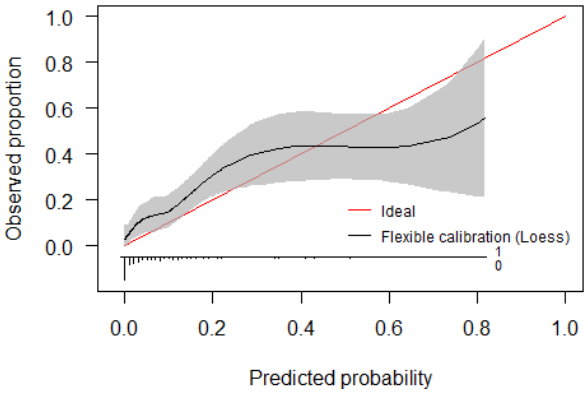
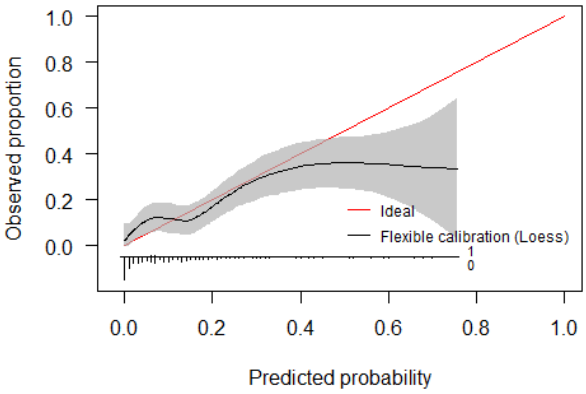


Figure A3 | Calibration curves of the dysphagia grade ≥ 2 models. (a) logistic regression, (b) ridge regression, (c) logistic regression after multiple imputation of the outcome with MICE, (d) ridge regression after multiple imputation of the outcome with MICE, (e) self-training with logistic regression as classifier, (f) self-training with ridge regression as classifier.

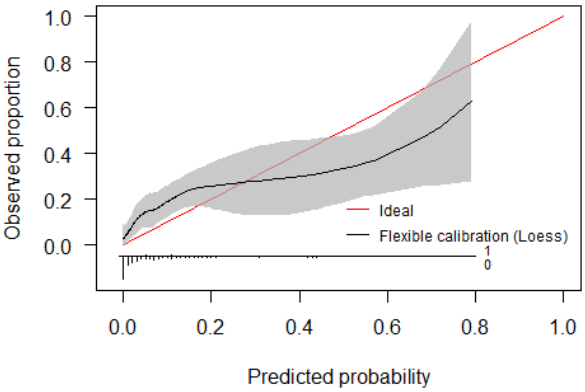
(a) Logistic regression



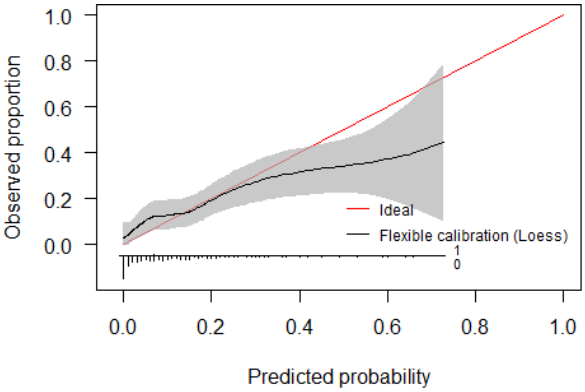
(b) Ridge regression



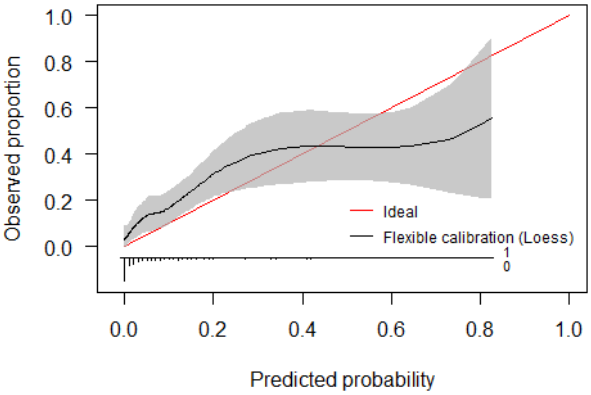
(c) Logistic regression after MICE



(d) Ridge regression after MICE



(e) Self-training with logistic regression as classifier



(f) Self-training with ridge regression as classifier

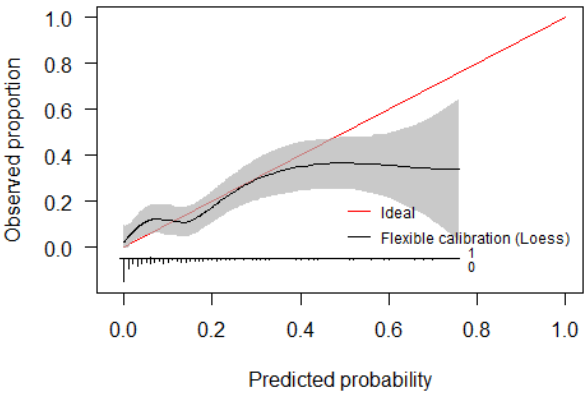


Figure A4 | Calibration curves of the dysphagia grade ≥ 3 models. (a) logistic regression, (b) ridge regression, (c) logistic regression after multiple imputation of the outcome with MICE, (d) ridge regression after multiple imputation of the outcome with MICE, (e) self-training with logistic regression as classifier, (f) self-training with ridge regression as classifier.

Table A1 | Regression coefficients of the models for xerostomia grade ≥ 2.

	(Intercept)	Mean dose to ipsilateral parotid (sqrt) + mean dose to contralateral parotid (sqrt)	Mean dose to submandibulars	Xerostomia at baseline grade ≥ 2	Xerostomia at baseline grade ≥ 3
Logistic regression	-2.605	0.157	0.011	0.502	1.078
Ridge regression	-2.425	0.121	0.015	0.44	0.965
MICE (logistic regression)	-2.62	0.166	0.008	0.542	1.226
MICE (ridge regression)	-2.438	0.124	0.014	0.473	1.099
Self-training (logistic regression)	-2.635	0.159	0.011	0.507	1.093
Self-training (ridge regression)	-2.446	0.121	0.015	0.444	0.968

Table A2 | Regression coefficients of the models for xerostomia grade ≥ 3.

	(Intercept)	Mean dose to ipsilateral parotid (sqrt) + mean dose to contralateral parotid (sqrt)	Mean dose to submandibulars	Xerostomia at baseline grade ≥ 2	Xerostomia at baseline grade ≥ 3
Logistic regression	-3.711	-0.019	0.033	0.037	1.191
Ridge regression	-3.383	0.042	0.016	-0.001	1.066
MICE (logistic regression)	-3.557	0.013	0.024	0.055	1.122
MICE (ridge regression)	-3.177	0.047	0.012	-0.002	0.944
Self-training (logistic regression)	-3.799	-0.021	0.034	0.054	1.238
Self-training (ridge regression)	-3.585	0.033	0.02	0.033	1.179

Table A3 | Regression coefficients of the models for dysphagia grade ≥ 2.

	(Intercept)	Mean dose to oral cavity	Mean dose to PCM superior	Mean dose to PCM medius	Mean dose to PCM inferior	Dysphagia at baseline grade ≥ 2	Dysphagia at baseline grade ≥ 3	Primary tumor location pharynx	Primary tumor location larynx
Logistic regression	-5.439	0.085	-0.005	0.003	0.019	0.914	1.028	-0.358	-0.279
Ridge regression	-3.906	0.036	0.015	0.011	0.005	0.957	1.146	-0.289	-0.532
MICE (logistic regression)	-5.204	0.082	-0.006	0.009	0.015	0.953	1.038	-0.515	-0.444
MICE (ridge regression)	-3.803	0.036	0.015	0.012	0.003	0.986	1.149	-0.347	-0.568
Self-training (logistic regression)	-5.588	0.088	-0.006	0.002	0.019	0.927	1.078	-0.318	-0.214
Self-training (ridge regression)	-3.946	0.036	0.016	0.011	0.005	0.965	1.166	-0.281	-0.529

Table A4 | Regression coefficients of the models for dysphagia grade ≥ 3.

	(Intercept)	Mean dose to oral cavity	Mean dose to PCM superior	Mean dose to PCM medius	Mean dose to PCM inferior	Dysphagia at baseline grade ≥ 2	Dysphagia at baseline grade ≥ 3	Primary tumor location pharynx	Primary tumor location larynx
Logistic regression	-10.61	0.094	-0.014	0.032	0.05	0.357	1.367	0.52	-0.216
Ridge regression	-7.684	0.037	0.016	0.021	0.036	0.506	1.532	0.25	-0.631
MICE (logistic regression)	-10.441	0.099	-0.013	0.033	0.045	0.231	1.093	0.248	-0.371
MICE (ridge regression)	-7.465	0.041	0.017	0.02	0.031	0.422	1.325	0.084	-0.681
Self-training (logistic regression)	-10.885	0.099	-0.015	0.028	0.053	0.354	1.409	0.616	-0.122
Self-training (ridge regression)	-7.8	0.038	0.016	0.021	0.036	0.525	1.561	0.28	-0.611

Table A5 | The total number of pseudolabeled observations (and the number of which are incorrect) and number of iterations in which new labels were added of the self-training method with logistic regression and ridge regression respectively, for each dataset with decreasing numbers of labelled observations of xerostomia grade ≥ 2.

	Self-training with logistic regression		Self-training with ridge regression	
	Pseudolabels (incorrect)	Iterations	Pseudolabels (incorrect)	Iterations
710 labelled / 40 unlabeled	5 (0)	1	4 (0)	1
670 labelled / 40 unlabeled	5 (0)	1	4 (0)	1
630 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
590 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
550 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
510 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
470 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
430 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
390 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
350 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
310 labelled / 40 unlabeled	5 (0)	2	4 (0)	1
270 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
230 labelled / 40 unlabeled	5 (0)	1	4 (0)	1
190 labelled / 40 unlabeled	4 (0)	1	4 (0)	1
150 labelled / 40 unlabeled	5 (0)	1	4 (0)	1
110 labelled / 40 unlabeled	5 (0)	1	4 (0)	3
70 labelled / 40 unlabeled	4 (0)	1	0 (0)	0

Table A6 | The total number of pseudolabeled observations (and the number of which are incorrect) and number of iterations in which new labels were added of the self-training method with logistic regression and ridge regression respectively, for each dataset with decreasing numbers of labelled observations of xerostomia grade ≥ 3.

	Self-training with logistic regression		Self-training with ridge regression	
	Pseudolabels (incorrect)	Iterations	Pseudolabels (incorrect)	Iterations
710 labelled / 40 unlabeled	32 (2)	1	32 (2)	1
670 labelled / 40 unlabeled	33 (3)	2	32 (2)	1
630 labelled / 40 unlabeled	32 (3)	2	32 (2)	1
590 labelled / 40 unlabeled	33 (3)	2	32 (2)	1
550 labelled / 40 unlabeled	31 (3)	2	32 (2)	1
510 labelled / 40 unlabeled	29 (3)	2	33 (3)	1
470 labelled / 40 unlabeled	30 (4)	2	34 (3)	2
430 labelled / 40 unlabeled	32 (5)	2	34 (3)	1
390 labelled / 40 unlabeled	31 (4)	3	33 (3)	2
350 labelled / 40 unlabeled	31 (4)	3	34 (3)	2
310 labelled / 40 unlabeled	29 (3)	2	33 (3)	2
270 labelled / 40 unlabeled	32 (4)	3	34 (3)	2
230 labelled / 40 unlabeled	30 (3)	1	32 (2)	1

Table A7 | The total number of pseudolabeled observations (and the number of which are incorrect) and number of iterations in which new labels were added of the self-training method with logistic regression and ridge regression respectively, for each dataset with decreasing numbers of labelled observations of dysphagia grade ≥ 2.

	Self-training with logistic regression		Self-training with ridge regression	
	Pseudolabels (incorrect)	Iterations	Pseudolabels (incorrect)	Iterations
710 labelled / 40 unlabeled	18 (1)	1	14 (0)	1
670 labelled / 40 unlabeled	19 (1)	1	14 (0)	1
630 labelled / 40 unlabeled	20 (1)	2	14 (0)	1
590 labelled / 40 unlabeled	20 (1)	2	14 (0)	1
550 labelled / 40 unlabeled	18 (1)	2	13 (0)	1
510 labelled / 40 unlabeled	18 (1)	1	14 (0)	1
470 labelled / 40 unlabeled	19 (1)	2	15 (0)	1
430 labelled / 40 unlabeled	20 (2)	2	16 (0)	1
390 labelled / 40 unlabeled	23 (4)	2	16 (0)	1
350 labelled / 40 unlabeled	20 (2)	2	17 (0)	1
310 labelled / 40 unlabeled	20 (2)	2	18 (1)	2
270 labelled / 40 unlabeled	20 (2)	2	16 (0)	1
230 labelled / 40 unlabeled	21 (2)	2	14 (0)	2

Table A8 | The total number of pseudolabeled observations (and the number of which are incorrect) and number of iterations in which new labels were added of the self-training method with logistic regression and ridge regression respectively, for each dataset with decreasing numbers of labelled observations of dysphagia grade ≥ 3.

	Self-training with logistic regression		Self-training with ridge regression	
	Pseudolabels (incorrect)	Iterations	Pseudolabels (incorrect)	Iterations
710 labelled / 40 unlabeled	29 (6)	1	22 (2)	1
670 labelled / 40 unlabeled	29 (6)	2	24 (3)	2
630 labelled / 40 unlabeled	29 (5)	2	25 (4)	2
590 labelled / 40 unlabeled	28 (6)	1	24 (3)	2
550 labelled / 40 unlabeled	29 (5)	2	24 (3)	1
510 labelled / 40 unlabeled	29 (5)	2	23 (2)	1
470 labelled / 40 unlabeled	28 (5)	2	24 (3)	1
430 labelled / 40 unlabeled	25 (3)	2	24 (3)	2
390 labelled / 40 unlabeled	25 (3)	2	24 (3)	1
350 labelled / 40 unlabeled	23 (3)	2	23 (3)	1
310 labelled / 40 unlabeled	22 (2)	2	23 (3)	1
270 labelled / 40 unlabeled	21 (2)	1	21 (2)	1
230 labelled / 40 unlabeled	21 (2)	1	22 (2)	1

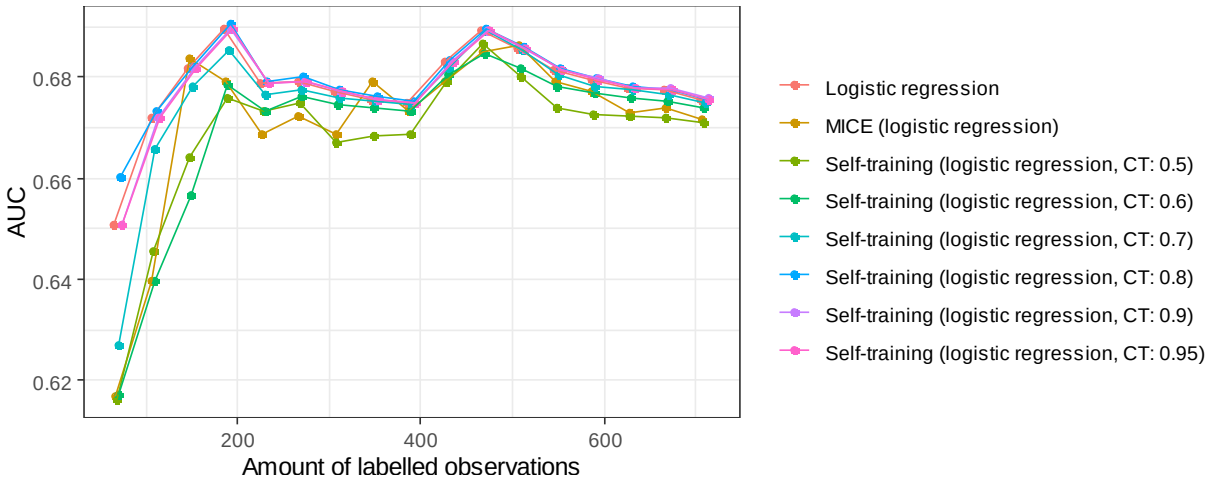
Table A9 | The total number of pseudolabeled observations (and the number of which are incorrect) and number of iterations in which new labels were added of the self-training method with logistic regression for different confidence thresholds. The methods were applied to datasets with decreasing numbers of labelled observations of xerostomia grade ≥ 2.

	Confidence threshold: 0.5		Confidence threshold: 0.6		Confidence threshold: 0.7	
	Pseudolabels (incorrect)	Iterations	Pseudolabels (incorrect)	Iterations	Pseudolabels (incorrect)	Iterations
710 labelled / 40 unlabeled	40 (16)	1	19 (4)	1	9 (1)	2
670 labelled / 40 unlabeled	40 (16)	1	19 (4)	1	8 (1)	1
630 labelled / 40 unlabeled	40 (17)	1	19 (4)	1	8 (1)	2
590 labelled / 40 unlabeled	40 (18)	1	19 (4)	1	8 (1)	1
550 labelled / 40 unlabeled	40 (17)	1	19 (4)	1	8 (1)	1
510 labelled / 40 unlabeled	40 (16)	1	19 (4)	1	6 (0)	2
470 labelled / 40 unlabeled	40 (14)	1	19 (4)	1	5 (0)	1
430 labelled / 40 unlabeled	40 (16)	1	19 (4)	1	8 (1)	1
390 labelled / 40 unlabeled	40 (18)	1	19 (4)	2	12 (1)	2
350 labelled / 40 unlabeled	40 (18)	1	21 (6)	3	12 (1)	2
310 labelled / 40 unlabeled	40 (18)	1	18 (4)	1	12 (1)	2
270 labelled / 40 unlabeled	40 (16)	1	19 (4)	1	12 (1)	2
230 labelled / 40 unlabeled	40 (15)	1	20 (4)	2	12 (1)	2
190 labelled / 40 unlabeled	40 (15)	1	19 (4)	2	9 (1)	2
150 labelled / 40 unlabeled	40 (14)	1	28 (9)	5	14 (2)	2
110 labelled / 40 unlabeled	40 (14)	1	34 (13)	5	16 (2)	4
70 labelled / 40 unlabeled	40 (17)	1	32 (13)	2	18 (5)	6

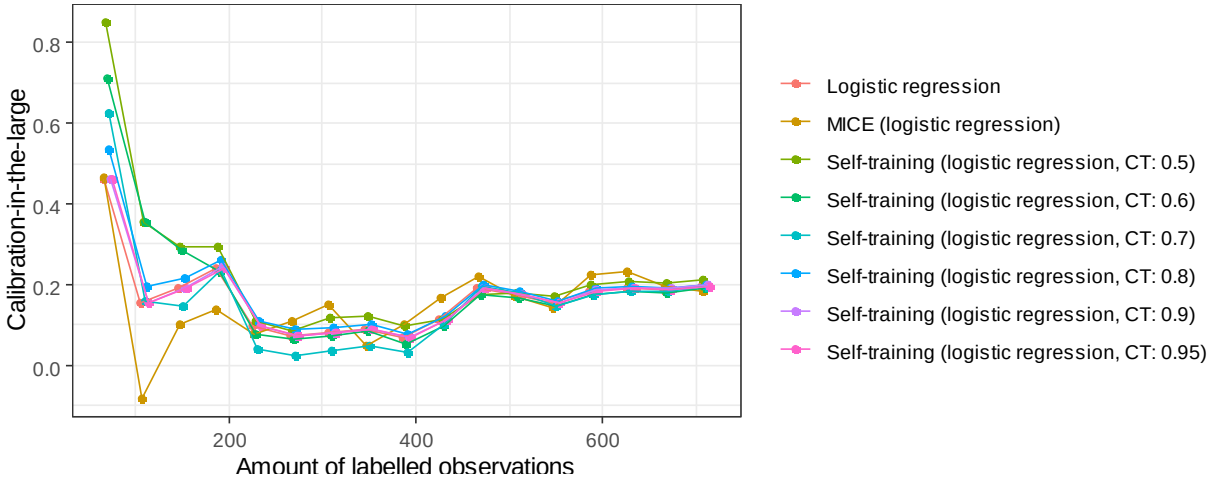
Table A9 continued

	Confidence threshold: 0.8		Confidence threshold: 0.9		Confidence threshold: 0.95	
	Pseudolabels (incorrect)	Iterations	Pseudolabels (incorrect)	Iterations	Pseudolabels (incorrect)	Iterations
710 labelled / 40 unlabeled	5 (0)	1	3 (0)	1	0	0
670 labelled / 40 unlabeled	5 (0)	1	3 (0)	2	0	0
630 labelled / 40 unlabeled	4 (0)	1	1 (0)	1	0	0
590 labelled / 40 unlabeled	4 (0)	1	1 (0)	1	0	0
550 labelled / 40 unlabeled	4 (0)	1	0	0	0	0
510 labelled / 40 unlabeled	4 (0)	1	0	0	0	0
470 labelled / 40 unlabeled	4 (0)	1	0	0	0	0
430 labelled / 40 unlabeled	4 (0)	1	0	0	0	0
390 labelled / 40 unlabeled	4 (0)	1	0	0	0	0
350 labelled / 40 unlabeled	4 (0)	1	0	0	0	0
310 labelled / 40 unlabeled	5 (0)	2	0	0	0	0
270 labelled / 40 unlabeled	4 (0)	1	0	0	0	0
230 labelled / 40 unlabeled	5 (0)	1	0	0	0	0
190 labelled / 40 unlabeled	4 (0)	1	1 (0)	1	0	0
150 labelled / 40 unlabeled	5 (0)	1	0	0	0	0
110 labelled / 40 unlabeled	5 (0)	1	0	0	0	0
70 labelled / 40 unlabeled	4 (0)	1	0	0	0	0

(a)



(b)



(c)

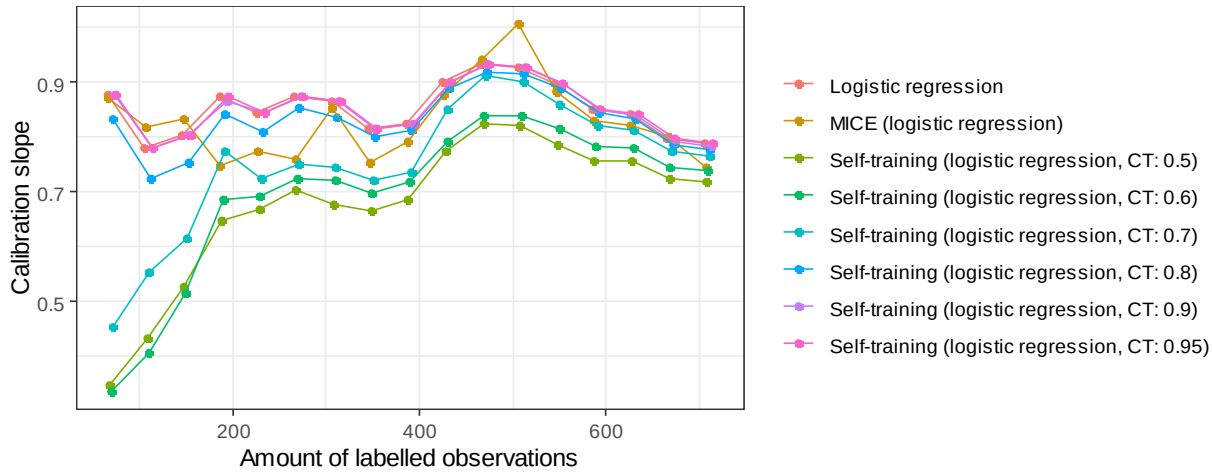


Figure A5 | External validation of the xerostomia grade ≥ 2 models with different confidence thresholds (CTs) for the self-training method with logistic regression. The x-axis shows the decrease in the amount of labelled data. The amount of unlabeled data is fixed at 40 observations. (a) The AUCs, (b) the calibration-in-the-large, and (c) the calibration slopes.

Table A10 | The total number of pseudolabeled observations (and the number of which are incorrect) and number of iterations in which new labels were added of the self-training method with logistic regression and ridge regression. The methods were applied to datasets with decreasing ratios of labelled vs unlabeled observations of xerostomia grade ≥ 2.

	Self-training with logistic regression		Self-training with ridge regression	
	Pseudolabels (incorrect)	Iterations	Pseudolabels (incorrect)	Iterations
710 labelled / 40 unlabeled	10 (0)	1	10 (0)	1
670 labelled / 80 unlabeled	15 (0)	1	14 (0)	1
630 labelled / 120 unlabeled	24 (0)	1	24 (0)	2
590 labelled / 160 unlabeled	37 (0)	2	36 (0)	2
550 labelled / 200 unlabeled	48 (3)	2	46 (2)	2
510 labelled / 240 unlabeled	58 (5)	2	56 (5)	3
470 labelled / 280 unlabeled	81 (8)	6	64 (4)	2
430 labelled / 320 unlabeled	102 (13)	5	95 (11)	7
390 labelled / 360 unlabeled	124 (19)	5	113 (15)	6
350 labelled / 400 unlabeled	130 (18)	4	121 (15)	7
310 labelled / 440 unlabeled	141 (21)	4	123 (16)	7
270 labelled / 480 unlabeled	149 (23)	5	138 (19)	4
230 labelled / 520 unlabeled	166 (27)	4	153 (21)	8
190 labelled / 560 unlabeled	203 (39)	7	174 (27)	5
150 labelled / 600 unlabeled	319 (79)	19	187 (30)	5

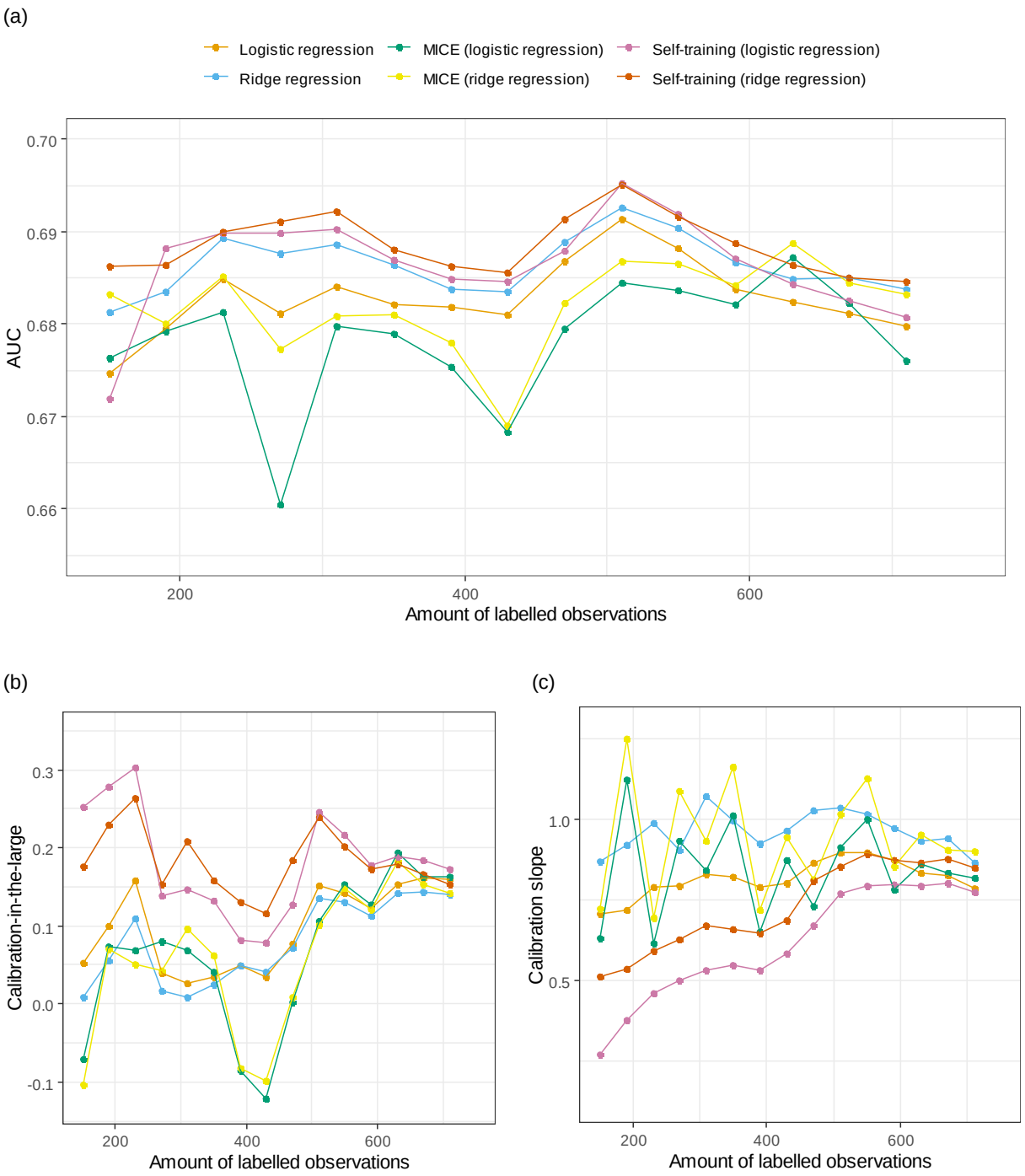


Figure A6 | External validation of the xerostomia grade ≥ 2 models with decreasing proportions of unlabeled data. The x-axis shows the amount of labelled data, while the total amount of observations (labelled plus unlabeled data) remains 750. (a) The AUCs, (b) the calibration-in-the-large, and (c) the calibration slopes.