Scientific Research Publishing

Open Journal of Statistics, 2018, 8, 497-512
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Adaptive Classification Methods for Predicting Transitions in the Nursing Workforce

George J. Knafl*, Mark Toles, Anna S. Beeber, Cheryl B. Jones

School of Nursing, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
Email: *gknafl@unc.edu

How to cite this paper: Knafl, G.J., Toles, M., Beeber, A.S. and Jones, C.B. (2018) Adaptive Classification Methods for Predicting Transitions in the Nursing Workforce. *Open Journal of Statistics*, **8**, 497-512.
https://doi.org/10.4236/ojs.2018.83032

Received: May 2, 2018
Accepted: June 10, 2018
Published: June 13, 2018

Abstract

Earlier analyses of transitions from licensed practical nurse (LPN) to registered nurse (RN) in the North Carolina (NC) nursing workforce in terms of 11 categorical predictors were limited by not considering parsimonious classifications based on these predictors and by substantial amounts of missing data. To address these issues, we formulated adaptive classification methods. Secondary analyses of data collected by the NC State Board of Nursing were also conducted to demonstrate adaptive classification methods by modeling the occurrence of LPN-to-RN transitions in the NC nursing workforce from 2001-2013. These methods combine levels (values) for one or more categorical predictors into parsimonious classifications. Missing values for a predictor are treated as one level for that predictor so that the complete data can be used in the analyses; the missing level is imputed by combining it with other levels of a predictor. An adaptive nested classification generated the best model for predicting an LPN-to-RN transition based on three predictors in order of importance: year of first LPN licensure, work setting at transition, and age at first LPN licensure. These results demonstrate that adaptive classification can identify effective and parsimonious classifications for predicting dichotomous outcomes such as the occurrence of an LPN-to-RN transition.

Keywords

Adaptive Classification, LPN-to-RN Transition, LPN Workforce

1. Introduction

In a previous analysis of nursing workforce data modeling the occurrence of a transition from a licensed practical nurse (LPN) to a registered nurse (RN) [1], it was anticipated that study findings would inform ongoing efforts to understand the supply and behaviors of the nursing workforce. Study findings were also in-

tended to describe potentially modifiable attributes of LPNs, who did and did not transition, that could be evaluated as opportunities for intervention. To achieve these aims, logistic regression analyses were conducted using 11 categorical characteristics as predictors, first generating unadjusted models one predictor at a time, and then generating a composite model using all 11 predictors in combination.

The data set was relatively large, with 37,781 observations. However, only two (18.2%) of the predictors had no missing values; missing values for the other nine predictors ranged from 2 (0.01%) to 7,041 (18.6%). The model based on all predictors used only 27,829 (71.0%) observations. Consequently, there was concern that missing data may have seriously affected study conclusions. Moreover, no attempt was made to remove extraneous terms from models, so generated models included non-significant terms (with p-values as large as 0.974 in unadjusted models and 0.958 in the composite model). To address these analysis issues, an exploratory approach was needed to systematically generate a parsimonious model using available categorical predictors while allowing for missing data and also accounting for the large sample size. Therefore, an adaptive classification approach addressing these issues was developed. This approach is presented here and demonstrated using the NC LPN workforce data.

Knafl and Ding [2] formulated and demonstrated adaptive regression methods for modeling nonlinear relationships for outcome (dependent, response) variables in terms of continuous predictor (independent, explanatory) variables. The adaptive regression modeling process is an analytic approach for conducting heuristic (*i.e.*, rule-based) searches through power transforms of primary predictors to generate an effective model for the data. Indicator variables (*i.e.*, dummy variables with values 0 or 1) as used to generate regression models equivalent to analysis of variance models can be considered in this search. For example, a categorical predictor C with three levels (or values) c_1, c_2, and c_3 can be represented by indicator variables I_1, I_2, and I_3 for $C = c_1$, $C = c_2$, and $C = c_3$, respectively. The adaptive regression search can consider any subset of these three indicator variables, but does not consider automatic adjustments to those indicator variables to address combinations of their underlying sets of observations such as the indicators $I_{1,2}$ for $C = c_1$ or $C = c_2$, $I_{1,3}$ for $C = c_1$ or $C = c_3$, and $I_{2,3}$ for $C = c_2$ or $C = c_3$. The adaptive classification approach presented here automatically considers such combinations. A missing value is treated as one of the levels for a categorical predictor so that the complete set of observations can be used in the adaptive classification. To avoid sparse classifications with the potential for over-fitting of the data, the adaptive classification process can be restricted to consider only classifications with all of their levels occurring for at least a specific percentage of the sample size. For example, reported analyses restricted the adaptive classification search to classifications with all levels occurring for at least 5% (or 1,890) of the LPN workforce data set.

Likelihood cross-validation (LCV) scores (defined later) are used to evaluate

and compare alternative models and to guide the adaptive classification process (as also used to guide the adaptive regression process). LCV scores for two models can be compared using LCV ratio tests, based on the χ^2 distribution [3] and so analogous to standard likelihood ratio tests. The significance level for these tests can be controlled. For example, reported analyses adjusted for the large sample size of the NC LPN workforce data by conservatively setting the significance level α to 0.001 rather than to the conventional value 0.05. The computation of LCV scores requires estimating model parameters on k randomly generated subsets of the data, and so computation time increases with the sample size and with the number k of subsets. For large sample sizes, computation times can be prohibitively long. However, LCV scores can be approximated for large enough sample sizes [4] by Akaike information criterion (AIC) scores [5], which can be used in such cases to reduce the computation time.

2. LPN Data

Data were collected annually by the NC State Board of Nursing from LPNs licensed in the state and maintained by the North Carolina Health Professions Data System at the Cecil G. Sheps Center for Health Services Research at the University of North Carolina at Chapel Hill. Data for the 2000-2013 LPN workforce were extracted from the Health Professions Data System.

Data were available for 37,781 LPNs licensed in the state of North Carolina (NC) from the years 2000 to 2013 with 3,161 (8.4%) of these experiencing an LPN-to-RN transition between 2001-2013 as indicated by the first time presence of an RN license number in the data set. Data were also available for 1,617 other LPNs who had made an RN transition prior to 2001; these data were not used in reported analyses.

A total of 11 categorical characteristics were available as potential predictors of an LPN-to-RN transition (**Table 1**). LPNs were primarily female (93.6%), White (69.4%), with a degree at first LPN licensure from a US school (95.1%), having a diploma as highest nursing degree (65.9%), working full time (64.0%), and residing in an urban location (73.0%). The categories for year and age at first LPN licensure were set to approximate quartile splits for the non-missing values [1]. The nine NC Area Health Education Centers (AHECs) are listed in **Table 1** in increasing size of the number of LPNs; this serves as a measure of geographical location within NC. The mission of these centers is to improve access to quality health care for the people of NC.

3. Data Analysis

3.1. LCV Scores

Models are evaluated and compared using k-fold likelihood cross-validation (LCV) scores. The data are first randomly partitioned into k distinct subsets, called folds in the statistics literature [6]. Fold likelihoods are calculated using the data in each fold and estimates of the parameter vector $\boldsymbol{\theta}$ computed with the

Table 1. Categorical characteristics for NC LPNs, 2000-2013.

Characteristic	Values	*N* (%)[a]
gender	female	35,356 (93.6)
	male	2,423 (6.4)
	missing	2 (0.01)
race/ethnicity	White	26,230 (69.4)
	Black	9,153 (24.2)
	American Indian	532 (1.4)
	Hispanic	538 (1.4)
	Asian	365 (1.0)
	other	651 (1.7)
	missing	312 (0.8)
year of first LPN licensure	1938-1981	8,515 (22.5)
	1982-1995	9,923 (26.3)
	1996-2004	9,017 (23.9)
	2005-2013	10,029 (26.5)
	missing	237 (0.8)
age at first LPN licensure (years)	16 - 22	9,182 (24.3)
	23 - 27	9,227 (34.4)
	28 - 34	9,527 (25.2)
	34 - 68	9,546 (15.3)
	missing	299 (0.8)
degree at first LPN licensure from US school	no	1609 (4.3)
	yes	35,948 (95.1)
	missing	224 (0.6)
highest nursing degree in last year as an LPN	diploma	24,889 (65.9)
	associate degree	3,688 (9.8)
	baccalaureate of science in nursing	555 (1.5)
	master of science in nursing	281 (0.7)
	doctorate in nursing	25 (0.1)
	missing	8,343 (22.1)
work setting in last year as an LPN	hospital in-patient	3,748 (9.9)
	long-term care	12,911 (34.2)
	solo/group practice or hospital out-patient	5,711 (15.1)
	other	8,458 (22.4)
	missing	6,953 (18.4)
specialty in last year as an LPN	community-based practice	3,412 (9.0)
	geriatrics	11,849 (31.4)
	medical/surgical	1,970 (5.2)
	pediatrics	2,175 (5.8)
	other	11,335 (30.0)
	missing	7,041 (18.5)
employed full time in last year as an LPN	no	6,777 (1.9)
	yes	24,176 (64.0)
	missing	6,828 (18.1)

Continued

located in rural area in last year as an LPN	no	24,579 (73.0)
	yes	10,202 (27.0)
located in NC AHEC in last year as an LPN	Area L	1,398 (3.7)
	South East	2,302 (6.0)
	Greensboro	3,497 (9.3)
	Mountain	4,120 (10.9)
	Southern Regional	4,744 (12.6)
	Eastern	4,819 (12.8)
	Charlotte	5,533 (14.6)
	Wake	5,555 (14.7)
	Northwest	5,813 (15.4)

AHEC—Area Health Education Center; LPN—licensed practical nurse; NC—North Carolina; US—United States. [a]Out of 37,781 LPNs.

rest of the data (hence the cross-validation). Fold likelihoods are multiplied together and the product is normalized by the sample size n (*i.e.*, by raising it to the power $1/n$) to generate LCV scores with larger values indicating better models. The same initial seed for starting the random number generation is used to randomly generate the folds for all models for the same data so that associated LCV scores are based on the same fold assignments and so are comparable.

Larger LCV scores do not necessarily indicate substantially (or distinctly or significantly) better models. This issue of a substantial improvement in the model can be addressed with LCV ratio tests. Let M denote a model for some data with n observations and M' a submodel with DF fewer parameters. DF is the associated degrees of freedom for a LCV ratio test between models M and M', and

$$D = 2\cdot\log(\text{LCV}(M)^n) - 2\cdot\log(\text{LCV}(M')^n)$$

is approximately χ^2 distributed with DF degrees of freedom (the power n is needed to remove the normalization of the LCV score by the power $1/n$). As for standard likelihood ratios, the log transform is required to produce an asymptotic χ^2 distributed statistic. This can be expressed in terms of the associated proportional decrease in the LCV score

$$\text{PD}(D,n) = \left(\text{LCV}(M) - \text{LCV}(M')\right)\Big/\text{LCV}(M) = 1 - \exp\left(-D/(2\cdot n)\right).$$

The proportional decrease PD(D, n) is substantial (or distinct or significant) if it exceeds the threshold PD($\delta(1-\alpha$, DF), n) where $\delta(1-\alpha$, DF) is the cutoff for a significant χ^2 test with DF degrees of freedom and significance level α. Equivalently, substantial improvements can be assessed using the percent decrease PD(D, n)·100% in place of the proportional decrease.

When the sample size is large, LCV scores can be approximated by AIC scores [4] formulated so that larger scores indicate better models and normalized by the sample size. Specifically, the usual smaller is better AIC score for model M with estimated parameter vector $\boldsymbol{\theta}$ is defined as

$$\mathrm{AIC}(M(\boldsymbol{\theta})) = -2\cdot\log(\mathrm{L}(\boldsymbol{\theta})) + 2\cdot\dim(\boldsymbol{\theta})$$

where $\mathrm{L}(\boldsymbol{\theta})$ is the likelihood for the data evaluated at the estimated parameter vector $\boldsymbol{\theta}$ and $\dim(\boldsymbol{\theta})$ is the dimension of that parameter vector equaling the number of model parameters. The associated adjusted AIC score is

$$\mathrm{AIC}^{+}(M(\boldsymbol{\theta})) = \exp(-\mathrm{AIC}(M(\boldsymbol{\theta}))/(2\cdot n)).$$

AIC^{+} ratio tests can be computed similarly to LCV ratio tests. Knafl and Ding ([2], pp. 68-69) demonstrate that AIC^{+} ratio tests are more conservative than standard likelihood ratio tests.

3.2. Adaptive Adjustment of an Individual Categorical Predictor *C*

Suppose that C is a categorical predictor with m nonmissing levels c_i having indexes i for $1 \le i \le m$. Starting with the full classification based on all m levels in their own groups and missing values if any in a separate group, systematically merge pairs of levels one at a time as follows. Compute LCV scores for each possible merger of two levels for the current classification. If the nonmissing levels of C are ordered (e.g., year at first LPN licensure levels), only consider mergers of consecutive nonmissing levels, for example, mergers of c_i with c_{i+1} for $1 \le i \le m-1$ at the first stage of the process. If the nonmissing levels of C are nominal (e.g., race), consider all pairs of two distinct levels for the current classification. If C has missing values, also consider all mergers of the missing level with each of the nonmissing levels. If the best LCV score for pairwise mergers of the current classification's levels generates a substantial percent decrease compared to the LCV score for the current classification, stop the search and use the current classification. The associated threshold for this LCV ratio test is based on DF = 1 because the number of levels has changed by 1. Otherwise (*i.e.*, when the percent decrease is not substantial) continue the adjustment process using the pairwise merger generating the best LCV score. Note that when C has missing values and the missing value level is merged with some other subset of nonmissing levels for C, the missing values have effectively been imputed as being one of the nonmissing levels in that subset.

3.3. Adaptive Additive Adjustment of a Classification Using a Second Categorical Predictor *C′*

Suppose that a classification based on a categorical predictor C has been adaptively generated using the above individual predictor adjustment process and that there are m^* levels corresponding to groupings of the m levels of C. Suppose that C' has m' nonmissing levels c'_i for $1 \le i \le m'$. Apply the above single categorical predictor adjustment process to the levels of C' to systematically merge them while also including the m^* levels of C in the model. These additive classification models are based on an intercept, a fixed set of indicator variables for $m^* - 1$ levels of the classification based on C and indicator variables for 1 less than the number of levels for the current classification based on C'.

Let $m^{\#}$ denote the number of levels for the current classification based on C'.

The additive model decomposes each of the m^* levels of the classification based on C into $m^\#$ levels for a total of $m^* \cdot m^\#$ cells corresponding to combinations of levels for the classifications based on C and C'. Consequently, additive adjustments can generate composite classifications with relatively large numbers of cells. Moreover, some of these cells can be sparse containing relatively small numbers of observations. Nested adjustments (as defined next) of the levels of C', that is, applied separately within each of the m^* levels of the classification based on C, can resolve these shortcomings.

3.4. Adaptive Nested Adjustment of a Classification Using a Second Categorical Predictor C'

Suppose that a classification based on a categorical predictor C has been adaptively generated using the above individual predictor adjustment process and that there are m^* levels corresponding to groupings of the m levels of C. Suppose that C' has m' nonmissing levels c'_i for $1 \le i \le m'$. Apply the above single categorical predictor adjustment process to the complete set of levels of C' nested within each of the m^* levels of the classification based on C. Compute LCV scores for each adjustment of a pair of levels of the current nested classification. If the best LCV score over all such nested adjustments generates a substantial percent decrease (using a LCV ratio test) compared to the LCV score for the current nested classification, stop the search and use the current nested classification. Otherwise continue the adjustment process considering further pairwise nested adjustments to the levels of the nested adjustment at the current stage of the process generating the best LCV score.

As an example, suppose the adaptive classification based on C has three levels c_1, c_2, and c_3 and C' has four nonmissing ordinal levels c'_1, c'_2, c'_3, and c'_4 and no missing values. The first stage of the nested classification considers the 3 pairwise ordered mergers of c'_1 with c'_2, c'_2 with c'_3, and c'_3 with c'_4 nested within each of the 3 levels c_1, c_2, and c_3 for a total of 9 pairwise mergers. The next nested classification is the one based on the pairwise merger of these 9 with the best LCV score, assuming that score is not substantially smaller (using a LCV ratio test) than the score for the classification based on only the levels c_1, c_2, and c_3. Suppose that this corresponds to the merger $c'_{1,2}$ of c'_1 with c'_2 nested within the level c_1. The next step in the process considers the same 6 pairwise mergers nested within the levels c_2 and c_3 as well as the 2 pairwise mergers of $c'_{1,2}$ with c'_3 and c'_3 with c'_4 nested within the level c_1. This nested adjustment process continues until the best LCV score for the next set of pairwise mergers generates a substantial percent decrease over the score for the current nested classification. If C' is nominal, than there are initially 6 possible mergers of distinct pairs of levels of C' within each of the 3 levels of the classification based on C for a total of 18 pairwise mergers. When C' also has missing values, extra mergers are considered pairing the missing level of C' with each of the 4 nonmissing levels of C' within each of the 3 levels of the classification based on C for a total of 12 extra pairwise mergers.

3.5. Handling More Than Two Categorical Predictors

The additive and nested adjustments defined above for adaptively combining a second categorical predictor with a classification previously adaptively generated from a first categorical predictor generalizes readily to adaptively combining one more categorical predictor with a classification adaptively generated from two or more other categorical predictors.

3.6. Adjusting an Adaptive Nested Classification

The adaptive nested classification process only considers nested adjustments within each combination of levels of a prior classification and not across those combinations of levels. If all categorical predictors are ordinal or all nominal, the final nested classification can be adjusted by recoding it as a single classification and applying the adaptive classification process to that recoded classification. If the LCV score increases, the classification has been improved; if it decreases but not substantially, the adjusted classification is a competitive, parsimonious alternative. An example is provided in Section 4.5.

3.7. Restricting the Search to Avoid Sparse Classifications

With a sparse classification defined as one with at least one level containing less than a fixed percentage of the n observations, continue any of the above adaptive classification searches if the current classification is sparse, even if that generates a substantial percent decrease in the LCV score. Once the current classification becomes nonsparse, it will remain that way throughout the rest of the process because levels increase in size with mergers. Apply the stopping rule for the search starting with the first nonsparse classification considered in the search.

3.8. Computation

All reported computations were conducted using SAS® version 9.4 (SAS Institute, Inc., Cary, NC). The adaptive classification process was implemented in a SAS macro. This macro and the SAS code used to generate the analyses are available at http://www.unc.edu/~gknafl/AdaptClass.html (accessed May 1, 2018).

4. Results

Reported adaptive classifications used the categorical characteristics of Table 1 to predict the occurrence of an LPN-to-RN transition. The significance level α was set at 0.001 to reflect the large sample size. Missing values were treated as an extra level so that all analyses used the complete data. Classifications were restricted to those with at least 5% (1,890) LPNs within each of their levels to avoid sparse classifications. The threshold for a substantial percent decrease in the LCV and AIC^+ scores for 37,781 observations, DF = 1, and significance level α = 0.001 was $PD(D, n) = 0.014\%$ with $D = \delta(0.999, 1) = 10.82757$ (this threshold value was generated in the output of the SAS adaptive classification macro).

Gender, race/ethnicity, work setting, specialty, and NC AHEC were nominal predictors; the other predictors were ordinal.

4.1. Comparison of LCV and AIC⁺ Scores

The adaptive classification of year of first LPN licensure was used to assess computation times and the approximation of LCV scores by the AIC⁺ score. LCV scores for $k = 5$, 10, and 15 folds were considered. For all three values of k and also for the AIC⁺ case, the single predictor adaptive classification process first merged the missing level with the 2005-2013 level and then stopped, generating the same 4-level classification. LCV scores rounded to five decimal digits were 0.76711, 0.76709, and 0.76708 for $k = 5$, 10, and 15 folds, respectively, compared to the AIC⁺ score of 0.76708. Clock times required for these computations increased from 23.6 minutes to 52.7 minutes, and then to 81.6 minutes for $k = 5$, 10, and 15 folds, respectively, compared to only 0.2 minutes for the AIC⁺ score.

Consequently, the adaptive classification of year of first LPN licensure was robust to the choice of score used to control the process. Also, the sample size was large enough to warrant use of the AIC⁺ score in place of LCV scores, which reduced the computation times to an acceptable level not possible with LCV scores. Consequently, only AIC⁺ scores were used in subsequent analyses.

4.2. Adaptive Classification of Individual Characteristics

Table 2 contains results for adaptive classification of the individual categorical characteristics of Table 1 for predicting an LPN-to-RN transition. Gender, race/ethnicity, and degree from a US school were not included in Table 2 because they generated constant classifications. Consequently, these three characteristics were not considered in subsequent analyses.

Odds ratios (ORs) for an LPN-to-RN transition are provided in Table 2. The reference categories were chosen so that all reported ORs are larger than 1, thereby indicating an increased chance of an LPN-to-RN transition. Confidence intervals and p-values were not reported for these ORs. Generated levels provided substantially different predictions of an LPN-to-RN transition due the adaptive classification heuristics; significance is thus a consequence of the analysis method and so seems inappropriate to report.

Missing values were not imputed for three characteristics: work setting, specialty, and employed full time. Imputation for the other six characteristics with missing values was primarily a result of restricting to nonsparse classifications with at least 5% of the LPNs in each level; of these six characteristics, only highest degree had more than 5% missing values (Table 1).

The adaptive classification based on the year of first LPN licensure generated the best (largest) AIC⁺ score, and so provided the best individual prediction of an LPN-to-RN transition. Consequently, this classification was used as the initial classification for multiple characteristics assessments, both additive and nested,

Table 2. Results of adaptive classification of individual categorical predictors of an LPN-to-RN transition from 2001-2013.

Characteristic	Values	OR[a]	AIC[+] Score
year of first LPN licensure	1938-1981	-	0.76708
	1982-1995	5.44	
	1996-2004	18.1	
	2005-2013 or missing	8.55	
age at first LPN licensure (years)	16 - 22	1.21	0.75111
	23-27 or missing	1.75	
	28 - 34	1.39	
	35 - 68	-	
highest nursing degree in last year as an LPN	diploma or associate degree	-	0.75106
	baccalaureate, master or doctorate in nursing or missing	1.54	
work setting in last year as an LPN	hospital in-patient	3.98	0.75651
	long-term care	1.49	
	solo/group practice or hospital out-patient or other	-	
	missing	2.47	
specialty in last year as an LPN	community-based practice	-	0.75413
	geriatrics, pediatrics, or other	1.44	
	medical/surgical	4.61	
	missing	2.36	
employed full time in last year as an LPN	yes	-	0.75120
	no	1.31	
	missing	1.67	
located in rural area in last year as an LPN	no	-	0.75035
	yes	1.29	

AHEC—Area Health Education Center; AIC[+]—adjusted Akaike Information Criterion; LPN—licensed practical nurse; NC—North Carolina; OR—odds ratio; RN—registered nurse. [a]OR for a LPN-to-RN transition relative to the category with OR setting "-".

of an LPN-to-RN transition. Only the other seven characteristics that generated nonconstant individual adaptive classifications were considered, and these adaptively reduced classifications were used in adaptive assessments rather than the original characteristics.

4.3. Adaptive Additive Classification of Multiple Characteristics

Additive adjustments based on five of the seven other characteristics generated the unadjusted classification based on year of first LPN licensure; most likely due to restricting to nonsparse classifications. The two exceptions corresponded to employed full time and located in a rural area. Employed full-time was adjusted

to the classification based on the employed full time level separate from the combined missing and employed part time levels. Located in a rural area was left unchanged.

The additive adjustment based on employed full time generated the larger AIC^+ score of 0.76818, which improved on the adaptive classification based on only year of first LPN licensure with AIC^+ score 0.76708 (Table 2). Further adaptive additive adjustment of this 2-characteristic additive classification using located in a rural area left the 2-characteristic additive classification unadjusted; thereby selecting that 2-characteristic classification as the final choice for additive adjustments.

Under this selected 2-characteristic classification, compared to year of first LPN license in 1938-1981, the OR for an increased chance of an LPN-to-RN transition was 18.4, 8.25, and 5.73 for the cases 1996-2004, 2005-2013 or missing, and 1982-1995, respectively. Also, compared to being employed full time, the OR for an increased chance of an LPN-to-RN transition was 1.51 for being employed part time or missing.

4.4. Adaptive Nested Classification of Multiple Characteristics

Table 3 presents results for the 2-characterisitic adaptive nested classifications. Results for located in NC AHEC were not included because it generated the unadjusted classification based on year of first LPN licensure by itself. The other six cases generated classifications nested within year of first LPN licensure. Age at first LPN licensure was nested in two or three levels within each level of year of first LPN licensure. The other five classifications only affected one or two of the year of first LPN licensure levels, leaving the other levels for year of first LPN licensure unadjusted.

Work setting nested within year of first LPN licensure generated the best AIC^+ score of 0.77028. This improved on the adaptive classification based on only year of first LPN licensure with AIC^+ score 0.76708 and on the best adaptive additive classification with AIC^+ score 0.76818. This 2-characteristic nested classification was used to generate 3-characteristic nested classifications based on the remaining five classifications that were nested within year of first LPN licensure.

The best 3-characteristic nested classification was based on further adjustments for age at first LPN licensure. All of the other four classifications had no effect, generating the model based on only work setting nested within year of first LPN licensure. Consequently, the adaptive nested classification search stopped with the final selected model described in Table 4. This model generated the best overall AIC^+ score of 0.77151. While it is based on three characteristics, there are only two levels of nesting with work setting nested within first year of LPN licensure 1996-2004 and with age at first LPN licensure nested within year of first LPN licensure 1982-1995. The other two years at first LPN licensure levels are unaffected by work setting and by age. This model has an acceptable *c*-index (or area under the receiver operating characteristics curve) of 0.72.

Table 3. Results of adaptive nested classification of two categorical predictors of an LPN-to-RN transition from 2001-2013 starting from the year of first LPN licensure.

Characteristic	Values	Year of First LPN Licensure	OR[a]	AIC[+] Score
age at first LPN licensure (years)	16 - 27 or missing	1938-1981	6.50	0.76954
	28 - 68		-	
	16 - 27 or missing	1982-1995	39.8	
	28 - 34		23.4	
	35 - 68		8.94	
	16 - 34 or missing	1996-2004	97.0	
	35 - 68		63.5	
	16 - 27 or missing	2005-2013 or missing	50.3	
	28 - 68		35.6	
highest nursing degree in last year as an LPN	-	1938-1981	-	0.76905
	-	1982-1995	5.44	
	diploma or associate degree	1996-2004	27.5	
	baccalaureate, master or doctorate in nursing or missing		14.1	
	diploma or associate degree	2005-2013 or missing	5.73	
	baccalaureate, master or doctorate in nursing or missing		10.4	
work setting in last year as an LPN	-	1938-1981	-	0.77028
	-	1982-1995	5.44	
	hospital in-patient or missing	1996-2004	33.5	
	long-term care or solo/group practice or hospital out-patient or other		12.1	
	-	2005-2013 or missing	8.55	
specialty in last year as an LPN	-	1938-1981	-	0.76903
	-	1982-1995	5.44	
	community-based practice or geriatrics, pediatrics, or other	1996-2004	13.8	
	medical/surgical or missing		31.3	
	-	2005-2013 or missing	8.55	
employed full time in last year as an LPN	-	1938-1981	-	0.76840
	-	1982-1995	5.44	
	yes	1996-2004	14.0	
	no or missing		27.0	

Continued

located in rural area in last year as an LPN	-	2005-2013 or missing	8.55	0.76737
	-	1938-1981	-	
	-	1982-1995	5.44	
	no	1996-2004	16.4	
	yes		22.9	
	-	2005-2013 or missing	8.55	

AIC⁺—adjusted Akaike Information Criterion; LPN—licensed practical nurse; NC—North Carolina; OR—odds ratio; RN—registered nurse. [a]OR for a LPN-to-RN transition relative to the category with OR setting "-".

Table 4. Final selected adaptive nested classification based on three categorical predictors of an LPN-to-RN transition from 2001-2013.

Age at First LPN Licensure	Work Setting in Last Year as an LPN	Year of First LPN Licensure	*N* (%)[a]	OR[b]	AIC⁺ Score
-	-	1938-1981	8,515 (22.5)	-	0.77151
16 - 34 years or missing	-	1982-1995	7,134 (18.9)	6.92	
35 - 68 years			2,789 (7.4)	1.84	
-	hospital in-patient or missing	1996-2004	2,893 (7.7)	33.5	
-	long-term care or solo/group practice or hospital out-patient or other		6,124 (16.2)	12.1	
-	-	2005-2013 or missing	10,326 (27.3)	8.55	

AIC⁺—adjusted Akaike Information Criterion; LPN—licensed practical nurse; OR—odds ratio; RN—registered nurse. [a]Out of 37,781 LPNs; [b] OR for an LPN-to-RN transition relative to the category with OR setting "-".

4.5. Example of an Adjusted Adaptive Nested Classification

The 2-characteristic nested classification between year of first LPN licensure and age at first LPN licensure is based on nine levels (Table 3). These two classifications are ordinal and so the nested classification can be considered to have nine ordinal levels. When this composite classification was treated as a single classification and subjected to the adaptive classification process, an eight level classification was generated, merging the seventh level based on the 1996-2004 year of first LPN licensure level and the 35 - 68 year at first LPN licensure level with the following or eighth level based on the 2005-2013 or missing year of first LPN licensure level and 16 - 27 or missing age at first LPN licensure level. The LCV score decreased from 0.76954 (Table 3) to 0.76947 with insubstantial percent decrease 0.009% (*i.e.*, less than the threshold of 0.014%), and so this was a competitive, parsimonious alternative classification. However, this did not affect the next stage of the adaptive nested classification process so that the final selected nested classification was still the one given in Table 4.

5. Discussion

Adaptive classification methods have been formulated and demonstrated using data on NC LPN-to-RN transitions in 2001-2013. These methods can be used to address individual categorical predictors as well as multiple categorical predictors combined through either additive or nested approaches. However, the additive approach can generate large numbers of combinations of levels for the individual predictors with sparse numbers within combinations. For the reported analyses, the nested approach produced models based on more individual predictors and with fewer combinations of levels than would be generated by the additive approach using the same number of predictors.

Adaptive classification can handle missing values without data loss by treating missing values as one more level for a categorical predictor, and in this case allowed for use of data from the complete NC LPN population (100% of the data compared to only 71% in earlier analyses). This approach is also used with multiple adaptive regression splines [7]. In reported analyses, models based on two or more characteristics had all missing levels for those characteristics combined with a nonmissing level, thereby imputing those missing values. This may not always happen, but the adaptive classification heuristics can be restricted to guarantee that all missing value levels be combined with some other level.

Under the selected best model (**Table 4**), compared to LPNs with first licensure in 1938-1981, the largest OR of 33.5 for an increased chance of an LPN-to-RN transition occurs for LPNs with first licensure in 1996-2004 and having hospital in-patient or missing work setting. The next largest OR of 12.1 occurs for LPNs with first licensure in 1996-2004 and having any other work setting, followed by an OR of 8.55 for LPNs with first licensure in 2005-2013 or missing, then an OR of 6.92 for LPNs with first licensure in 1982-1995 and age at first LPN licensure 16 - 34 years or missing, and finally by an OR of 1.84 for LPNs with first licensure in 1982-1995 and age at first LPN licensure 35 - 68 years.

These adaptive classification results suggest some important conclusions about the LPN workforce data. The chance of an LPN-to-RN transition can be reasonably treated as depending entirely on three characteristics: year at first LPN licensure, work setting in last year as an LPN, and age at first LPN licensure, in order of importance. While this chance also depends individually on other characteristics (**Table 2**), these individual dependencies are reasonable considered to be explainable by the joint dependence on the three primary characteristics. The results of earlier analyses [1] suggest the opposite conclusion that characteristics other than these three are also of importance, but that conclusion was based on data for only 71% of the LPNs and using all available characteristics without attempting to identify a parsimonious alternative model.

The odds of a transition change with cohort based on the year of first LPN licensure (**Table 2**) varies from 5.44 to 18.1 compared to the earliest 1938-1981 cohort with the lowest chance of a transition. However, these cohort effects be-

come more complex after further consideration of work setting and age (Table 4). For the 1982-1995 cohort, the odds for a transition interact with the age at first LPN licensure and are stronger for younger (OR = 6.92) than older ages (OR = 1.84). For the 1996-2004 cohort, the odds for a transition interact with the work setting in the last year as an LPN and are stronger for the hospital in-patient setting (OR = 33.5) than for other settings (OR = 12.1). For nurses in the other two cohorts, the chance of a transition is unaffected by any of the other classifications. These findings support the hypothesis that the odds of LPN-to-RN transitions are more prevalent in later cohorts than the 1938-1981 cohort, but not monotonically increasing as cohorts get more recent. Moreover, the largest odds of a transition occurred for the 1996-2004 cohort when working in a hospital in-patient setting and the next largest for the other LPNs in this cohort. Future studies designed to recruit LPNs for RN training and advancement might seek out nurses in the more recent cohorts, especially those employed in hospital in-patient, long-term care, and primary care settings.

A critical policy objective for the nursing workforce literature is the imperative to develop and sustain a sufficient supply of quality RNs. An important policy implication for the reported results is that a substantially greater effort will be needed to facilitate LPN-to-RN transitions in the workforce. These professional transitions were uncommon, occurring for only 8.4% of the NC LPNs over a 13 year period. Evidence from this study suggests that differences in age of LPN licensure and work setting have, for different cohorts, influenced LPN-to-RN transitions. Future investigators and policy makers might intervene to locate nurses in these groups and stimulate their interest in pursuing opportunities for career advancement. Qualitative research is needed to identify modifiable barriers to an LPN-to-RN transition and to better understand strategies that foster them.

6. Limitations

Adaptive classification methods are not directly supported by standard statistical software. However, a SAS macro has been developed to support those methods whose use only requires relatively basic knowledge of the SAS system. The formulation of these methods is based on an agglomerative or bottom-up approach combining larger sets of levels into smaller sets. A divisive or top-down approach is also possible, decomposing smaller sets of levels into larger sets, but has not yet been implemented. Reported analyses only addressed the logistic regression case with a dichotomous outcome, but the methods generalize to the ordinal regression case with an ordinal outcome with more than two outcome values, the multinomial regression case with a nominal outcome with more than two values, the linear regression case with a continuous outcome, and the Poisson regression case with a count outcome. Reported analyses also only addressed the univariate outcome case, but the methods generalize to the multivariate outcome case. Other methods could have been used instead, for example, multiple

adaptive regression splines [7] or classification and regression trees [8]. Additional, unexamined factors are likely also to contribute to the occurrence of LPN-to RN-transitions, such as cultural, financial, pedagogical, and systemic educational barriers that may impede career advancement in this population. Further work is needed to investigate the impact of these other factors.

7. Conclusions

Reported analyses demonstrate that adaptive classification can identify effective and parsimonious classifications for predicting dichotomous outcomes such as the occurrence of an LPN-to-RN transition. Moreover, these methods provide novel and meaningful insights that can inform policy making and workforce planning. These methods can be used more broadly, for example, other kinds of transitions not only in the workforce context but also for patient transitions or transitions of any kind.

Acknowledgements

This project was funded through HRSA Cooperative Agreement U881HP26495: Health Workforce Research Centers Program.

References

[1] Jones, C.B., Toles, M., Knafl, G.J. and Beeber, A.S. (2018) An Untapped Resource in the Nursing Workforce: Licensed Practical Nurses Who Transition to Become Registered Nurses. *Nursing Outlook*, **66**, 46-55. https://doi.org/10.1016/j.outlook.2017.07.007

[2] Knafl, G.J. and Ding, K. (2016) Adaptive Regression for Modeling Nonlinear Relationships. Springer International Publishing, Switzerland. https://doi.org/10.1007/978-3-319-33946-7

[3] Stone, M. (1977) An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society Series B*, **39**, 44-47.

[4] Claeskens, G. and Hjort, N.L. (2009) Model Selection and Model Averaging. Cambridge University Press, Cambridge.

[5] Sclove, S.L. (1987) Application of Model-Selection Criteria to Some Problems in Multivariate Analysis. *Psychometrika*, **52**, 333-343. https://doi.org/10.1007/BF02294360

[6] Burman, P. (1989) A Comparative Study of Ordinary Cross-Validation, ν-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika*, **76**, 503-514. https://doi.org/10.1093/biomet/76.3.503

[7] Friedman, J.H. (1991) Multivariate Adaptive Regression Splines. *Annals of Statistics*, **19**, 1-67. https://doi.org/10.1214/aos/1176347963

[8] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1998) Classification and Regression Trees. CRC Press, Boca Raton, FL.