Rural and Remote Health rrh.org.au
James Cook University ISSN 1445-6354

ORIGINAL RESEARCH

Using machine learning to identify factors associated with practice location of the healthcare workforce

AUTHORS


Jerry Bounsanga[1] MStat, Statistician


Martin S Lipsky[2] MD, Faculty


Eric S. Hon[3] AB, Manager


Frank W Licari[4] DDS, Dean


Clark Ruttinger[5] MBA, Research Director


Andrew Salt[6] BS, Research Specialist


Man Hung[7] PhD, Research Dean *

CORRESPONDENCE

*Dr Man Hung mhung@roseman.edu

AFFILIATIONS

[1] Quality Outcomes Research and Assessment, School of Medicine, University of Utah Health, Salt Lake City, UT 84108, USA

[2] Roseman University of Health Sciences, South Jordan, UT 84095, USA

[3] Department of Economics, University of Chicago, Chicago, IL 60637, USA

[4, 7] College of Dental Medicine, Roseman University of Health Sciences, South Jordan, UT 84095, USA

[5, 6] Utah Medical Education Council, Salt Lake City, UT 84102, USA

PUBLISHED

4 February 2022 Volume 22 Issue 1

HISTORY

RECEIVED: 29 July 2021

ACCEPTED: 12 October 2021

CITATION

Bounsanga J, Lipsky MS, Hon ES, Licari FW, Ruttinger C, Salt A, Hung M. Using machine learning to identify factors associated with practice location of the healthcare workforce. Rural and Remote Health 2022; 22: 7050. https://doi.org/10.22605/RRH7050

ABSTRACT:

Introduction: Past studies examined factors associated with rural practice, but none employed newer machine learning (ML) methods to explore potential predictors. The primary aim of this study was to identify factors related to practice in a rural area. Secondary aims were to capture a more precise understanding of the demographic characteristics of the healthcare professions workforce in Utah (USA) and to assess the viability of ML as a predictive tool.

Methods: This study incorporated four datasets – the 2017 dental workforce, the 2016 physician workforce, the 2014 nursing workforce and the 2017 pharmacy workforce – collected by the Utah Medical Education Council. Supervised ML techniques were used to identify factors associated with practice location, the outcome variable of interest.

Results: The study sample consisted of 11 259 healthcare professionals with an average age of 46.6 years, of which 36.6% were males and 94.5% Caucasian. Four ML methods were applied to assess model performance by comparing accuracy, sensitivity, specificity and area under the receiver operating characteristic (ROC) curve. Of the methods used, support vector machine performed the best (accuracy 99.7%, precision 100%, sensitivity 100%, specificity 99.4% and ROC 0.997). The models identified income and rural upbringing as the top factors associated with rural practice.

Conclusion: By far, income emerged as the most important factor associated with rural practice, suggesting that attractive income offers might help rural communities address health professional shortages. Rural upbringing was the next most important predictive factor, validating and updating earlier research. The performance of the ML algorithms suggests their usefulness as a tool to model other databases for individualized prediction.

Keywords:

dental, health care, location, machine learning, nursing, USA, workforce, pharmacy, physician.

FULL ARTICLE:

Introduction

Rural communities experience poorer health outcomes, lower life expectancies and more chronic disease and hospitalizations than communities in more densely populated urban environments[1]. Factors contributing to rural health disparities include older populations, lower incomes, less education, fewer occupational opportunities and less healthy lifestyle behaviors[2,3]. Additionally, rural areas struggle to both attract and retain physicians, nurses, dentists and pharmacists, creating access barriers[4] that further rural health disparities[5]. These shortages make research that identifies factors related to health professionals selecting careers in rural settings a topic of interest to public health officials, policy-makers, training institutions and others interested in rural health.

Previous research addressing the rural healthcare professional workforce primarily focused on rural training and developing a pipeline of health professionals interested in rural practice. These studies identified rural clinical rotations as a strong predictor for choosing rural medicine[6-11]. Other factors identified as influencing physician retention were rural upbringing and financial incentives such as loan forgiveness[11,12]. Factors such as race and size of undergraduate college were not associated with rural medicine practice[12]. Some studies found that men were more likely to practise rural medicine than women[9,13] while others did not find a link between being male and rural practice[12]. One study connected personality traits to rural practice and found physicians scoring higher on openness to experience, agreeableness and self-confidence more likely to choose rural medicine[14].

Research about other healthcare professions, such as dentists, nurses and pharmacists, also linked an individual's rural upbringing and rural training experiences to choosing a rural practice setting[15-18]. Dentists were more likely to practise in rural areas if they were male or had positive experiences with rural dental role models[16,19]. An Iowan study found that a dentist's birth state was not associated with rural practice decisions, but observed a correlation with being older and being female in solo practice[20]. Rural nurses tended to have less nursing education and were more likely to work full time in public/community health, long-term care, or ambulatory care settings than their non-rural counterparts[21]. Other studies addressing the rural nursing workforce centered on job satisfaction, turnover rates and burnout[22,23] and less on factors influencing career choice. Studies related to the pharmacy workforce are limited, but report that rural pharmacists are more likely to have rural roots and have received some type of rural exposure in their clinical training[17,24].

Despite the existing body of research examining rural workforce predictors, a systematic review by Lee and Nichols[8] concluded there is a need for more rigorous analysis and research related to implementing rural recruitment and retention strategies. More recently, Grobler et al[25] conducted an extensive systematic literature review to reconcile and update the literature on the healthcare professions workforce to help determine effective incentives to retention. They concluded that many previous studies

were limited by bias and confounding factors and expressed the need for more well-designed studies related to factors associated with choosing rural practice. Moreover, they cited a survey study by Trickett-Shockey et al[26] that challenged previous findings identifying rural upbringing or rural training and education as a strong predictor of student intent to choose rural practice. Though small, this study illustrates the complexity of predicting professional practice settings and suggests that choice results from a nexus of underlying factors.

Another limitation to the published literature is that earlier researchers primarily used traditional statistical methods. To the authors' knowledge, no existing study has applied newer, more robust analytics such as machine learning (ML). The algorithms employed by ML are known to be valuable, practical and applicable to a wide range of research questions, especially in health care[27]. ML can use data to detect patterns to predict outcomes, and advanced analytics from ML could improve the accuracy and precision of rural modelling and prediction as well as validate earlier findings[3]. The technique is useful to identify relationships between multiple data inputs or 'features' and an outcome. In ML, the computer learns by testing multiple sets of algorithms on a training dataset to determine which data variables help to classify an outcome. The results from using new analytic techniques should be of interest to educators, policy-makers and others interested in rural health. ML methods can also stimulate hypothesis testing research to explore and test previously identified associations.

This study seeks to add to the understanding of the factors related to rural practice. Specifically, it will be the first to apply ML techniques to a database and to assess the utility of ML as a tool to identify factors predicting the decision to practice in a rural area. Because the study uses a regional database, a secondary aim is to capture a more defined understanding of the demographic characteristics, of the Utah (USA) healthcare professions workforce.

Methods

Study design

This study used data collected by the Utah Medical Education Council (UMEC). Utah is a state in the Western Mountain region of USA with a population of about 3.2 million, of which 335 000 live in rural areas. UMEC gathers data on the supply of healthcare professions on different cycles every calendar year and includes information on demographic characteristics, practice settings, education, hours worked per week and career outlook. Each discipline reported on income by selecting from a range of income classifications. For example, physicians selected among 12 salary classifications before taxes and excluding benefits ranging from from less than US$49,000 to more than US$300,000 and after excluding residents and fellows yielded a median hour adjusted income for primary care providers of US$178,000 and US$229,000 for specialists[28]. Similarly, the other disciplines allowed respondents to select an income classification but used classifications that reflected the salary ranges of the different health professions. Most data were collected through paper surveys and Qualtrics, an online surveying tool. All data were de-identified before running any analyses.

Data processing

Datasets from four healthcare professions were cleaned and merged to run analyses. These four unpublished UMEC datasets include the 2017 dental workforce, the 2016 physician workforce, the 2014 nursing workforce and the 2017 pharmacy workforce. Similar variables for each dataset were identified and recoded to match accordingly. For example, data for practice settings in the nursing workforce had different values (ie 2=hospitals) compared to the physician workforce (ie 1=hospitals) and needed to be matched in order for values to represent the same practice settings.

Measures

Primary practice location was the outcome variable of interest for this study. For the purpose of this study, the variable was categorized as practising either in a non-rural (urban or suburban area) or in a rural area and was computed by taking the primary zip code and converting it to rural and non-rural areas based on rural-urban commuting area (RUCA) codes. Although there are different ways to classify rural and non-rural areas, RUCA codes, which were developed by the United States Department of Agriculture Economic Research Service, are widely used in healthcare research[29-31]. RUCA codes take into account census tracts based on daily commuting patterns, population density and urbanization, and are split into 33 distinct categories. For this study, these categories were clustered into two simple levels: non-rural (RUCA codes 1, 2 and 3, representing communities with ≥50 000 residents) and rural (RUCA codes 4 and above, or ≤49 999 residents).

Feature selection

Data included 20 149 cases and 88 variables (ie features or attributes) obtained by merging the dental, physician, nursing and pharmacy workforces. Surveys were sent to individuals working in Utah who held active Utah licenses for dentistry, medicine, nursing or pharmacy. Response rates for the survey were 50.8% for dentistry, 47% for physicians, 42% for nursing and 30.4% for pharmacy. As mentioned in the workforce reports from UMEC, these response rates were considered satisfactory due to meeting a sufficient 95% confidence interval. In addition to general questions that crossed disciplines, each survey contained questions relevant to the specific health profession. Further information regarding the UMEC workforce supply surveys has been published previously[32-34]. In preparation for data analyses, this study excluded non-relevant variables (such as license ID) and variables with more than 50% of missing data among the survey respondents; this yielded a total sample size of 11 259 (dental n=902, physician n=2587, nursing n=6932, pharmacy n=838) for outcome prediction of practice location. Variables with more than 50% of missing data were excluded to ensure reliable precision and accuracy while limiting errors that may occur with the ML algorithms. The random forest (RF) method was then applied to

select the most important variables or features for outcome prediction, known as feature selection. The RF method takes data inputs and randomly applies them to multiple decision trees iteratively until it identifies data features that help predict an outcome. In contrast to traditional statistical models, the models created by ML algorithms are extremely complex. Thousands of rules or parameters might be tested to define the model and consequently the exact internal processing pathways may be hard to identify. In this study, the RF method isolated 28 features for inclusion in the next steps of model building and validation. Feature selection helps prevent over-fitting of data and reduces errors in model complexity along with training time[35]. Appendix I outlines which features among the 28 variables, such as gender, race and debt, were dropped.

Analytical methods

The outcome variable of interest was practice location. To adjust for an observed imbalance of data, the analyses employed a replacement strategy to create a more balanced dataset to adjust for the minority class (eg the rural class of practice location, which had a much smaller sample size). Without adjusting for the minority class of data, ML methods often fail to correctly predict the minority class and cause an inflated model performance to the majority class. Oversampling, also known as sampling with replacement, has been used in previous ML studies because it is effective in treating class imbalance with large datasets[36-39]. Adjustment using sampling with replacement can reduce gaps between sensitivity, specificity and errors of misclassification[36,40]. Thus, the minority class in the data was adjusted (proportions of the minority class were resampled until reaching a similar sample size to the majority class), resulting in a more balanced dataset of 20 291 cases with 10 130 cases classified as non-rural and 10 161 cases classified as rural for modelling.

This study used several different supervised ML methods including decision tree (DT), RF regression, extreme gradient boosting (XGBoost) and support vector machine (SVM) for modelling. In supervised ML methods, the outcome of the study is known, predetermined, or preset by the data scientist or researcher. For example, the outcome of this study was preset to be the practice site. However, in unsupervised ML, the outcome is determined by the machine during the course of data exploration, making supervised ML methods more suitable for prediction studies and unsupervised ML methods more appropriate for studies focusing on clustering and feature reduction. In contrast to more traditional statistical methods such as logistic regression, ML includes higher-order interactions and examines complex non-linear relationships between model variables and outcomes. The ML methods in this study were chosen because of their established applicability in healthcare research, capability of over-fitting prevention, simplicity of comprehension, and general acceptance as useful ML methods[41-45]. Application of ML involves both training and test datasets, where algorithms applied to a training dataset can help identify associations that might be challenging to observe in complex and larger datasets. After the training dataset explores and ultimately predicts the outcome variable, the prediction is validated by comparing it against the test dataset, recognized as the validation set. The model that performs the best through the stages of validation will be the final chosen model. In this study, the data were randomly split into training and testing sets for model building and validation, using DT, RF, XGBoost and SVM. An 80/20 split (ie 80% of data for training and 20% of data for validation) was chosen based on previous literature and is referred to as the 80/20 rule or the Pareto principle[44]. More specifically, 80% of data were trained using k-folds cross validation and then 20% of data were tested for validation to minimize issues of over-fitting and model errors[46-48]. The metrics used to evaluate model performance included accuracy, sensitivity, specificity and area under the curve (AUC) of the receiver operating characteristic (ROC). Several sources provide detailed descriptions of ML and the techniques used in this study[49-51].

Descriptive statistics on demographic characteristics and clinical practice were analyzed for all the healthcare professions. The ML analyses were conducted using WEKA v3.9.4 (https://www.cs.waikato.ac.nz/ml/weka). Other statistical analyses, such as descriptive statistics were performed using SPSS v25.0 for Windows (IBM; http://www.spss.com).

Ethics approval

Roseman University of Health Science Institutional Review Board conducted ethical approval of this study and determined this study as non-human subject research.

Results

The study sample consisted of 11 259 healthcare professionals licensed in Utah, of which 36.6% were male and 63.4% were female with an average age of 46.6 years (standard deviation (SD) 12.98). Of the sample group, most healthcare professionals were Caucasian (n=10 375, 94.5%), went to a school outside of Utah (n=7024, 62.4%), and had a non-rural upbringing (n=8128, 73.5%). Only 1.9% (n=318) identified as being Hispanic. Table 1 summarizes the sample group demographics. There were significant differences of profession (p<0.001), race (p<0.001) and upbringing (p<0.001) between rural and non-rural practice location (see Table 1).

The average age of the rural health professions workforce was 46.6 years (SD=12.95), and 48% (n=538) of them worked in hospital settings (Table 2). The top five specialties for rural practice were general surgery (n=187, 17.0%), primary care consisting of general internal medicine, pediatrics and family medicine (n=177, 16.1%), other specialty (n=157, 14.2%), general dentistry (n=91, 8.3%) and emergency medicine (n=85, 7.7%).

The non-rural healthcare workforce had an average age of 46.6 years (SD=12.99) and about half worked in hospital settings (n=5096, 50.7%). The top five specialties for non-rural practice were other specialty (n=1564, 15.4%), primary care defined as general internal medicine, pediatrics and family medicine (n=1458, 14.4%), general surgery (n=946, 9.3%), general pharmacist (n=673, 6.6%) and critical care medicine (n=646, 6.4%).

Among the ML methods, the best performing classifier was SVM (accuracy 99.7%, precision 100%, sensitivity 100%, specificity 99.4%), followed by XGBoost (accuracy 96.6%, precision 100%, sensitivity 93.1%, specificity 100%), RF regression (accuracy 96.6%, precision 93.7%, sensitivity 100%, specificity 93.2%) and DT (accuracy 89.0%, precision 83.4%, sensitivity 97.5%, specificity 86.0%) (Table 3). Figure 1 presents the feature importance graph for SVM, and includes the 10 most important predictors: income, upbringing, total hours, age, years until retirement, school state, patients per week, degree year, practice setting and specialty. Importance scores are derived by constructing a prediction model in which variables that influence the model the most have the greatest impact on reducing model error. When variables with high *r* importance are excluded from the prediction model, increased model error occurs[38,52,53]. The higher the scores for these features, the more important they were in identifying rural practice location. In terms of rural practice decisions, income and upbringing were found to be the most important features. The ROC curves indicated that all ML algorithms performed exceptionally well, where curves placed closer to the top-left corners represent better performance (Fig2). Table 3 lists the model performance evaluations. Table 2 shows the descriptive statistics of the top 10 important features derived from SVM.

Table 1: Demographic characteristics of the healthcare professions workforce

Variable	Rural (*n*=1129)	Non-rural (*n*=10 130)	*p*-value	Overall (*n*=11 259)
Age (years)			0.928	
Mean±SD	46.59±12.95	46.63±12.97		46.63±12.98
Median (range)	46.00 (22–87)	46.00 (21–88)		46.00 (21–88)
Gender, *n* (%)†			0.213	
Male	392 (34.9)	3714 (36.8)		4106 (36.6)
Female	730 (65.1)	6371 (62.9)		7101 (63.4)
Profession, *n* (%)			<0.001	
Dentist	109 (9.7)	793 (7.8)		902 (8.0)
Physician	191 (16.9)	2396 (23.7)		2587 (23.0)
Registered nurse	743 (65.8)	6189 (61.1)		6932 (61.6)
Pharmacist	86 (7.6)	752 (7.4)		838 (7.4)
Race, *n* (%)†			<0.001	
American Indian	15 (1.4)	35 (0.4)		50 (0.5)
African American	4 (0.4)	51 (0.5)		55 (0.5)
Asian	15 (1.4)	338 (3.4)		353 (3.1)
Caucasian	1066 (95.9)	9309 (94.3)		10 375 (92.1)
Polynesian	2 (0.2)	43 (0.4)		45 (0.4)
Other	9 (0.8)	95 (1.0)		104 (0.9)
Ethnicity, *n* (%)†			0.053	
Non-Hispanic	1108 (98.8)	9880 (98.0)		10 988 (98.1)
Hispanic	13 (1.2)	201 (2.0)		214 (1.9)
Upbringing, *n* (%)†			<0.001	
Rural	705 (63.9)	2220 (22.3)		2925 (26.5)
Non-rural	399 (36.2)	7729 (77.7)		8128 (73.5)
School attended, *n* (%)			0.088	
Outside of Utah	678 (60.1)	6346 (62.6)		7024 (62.4)
Within Utah	451 (39.9)	3784 (37.4)		4235 (37.6)

† The sample size does not equal the total *N* because of non-response.
SD, standard deviation.

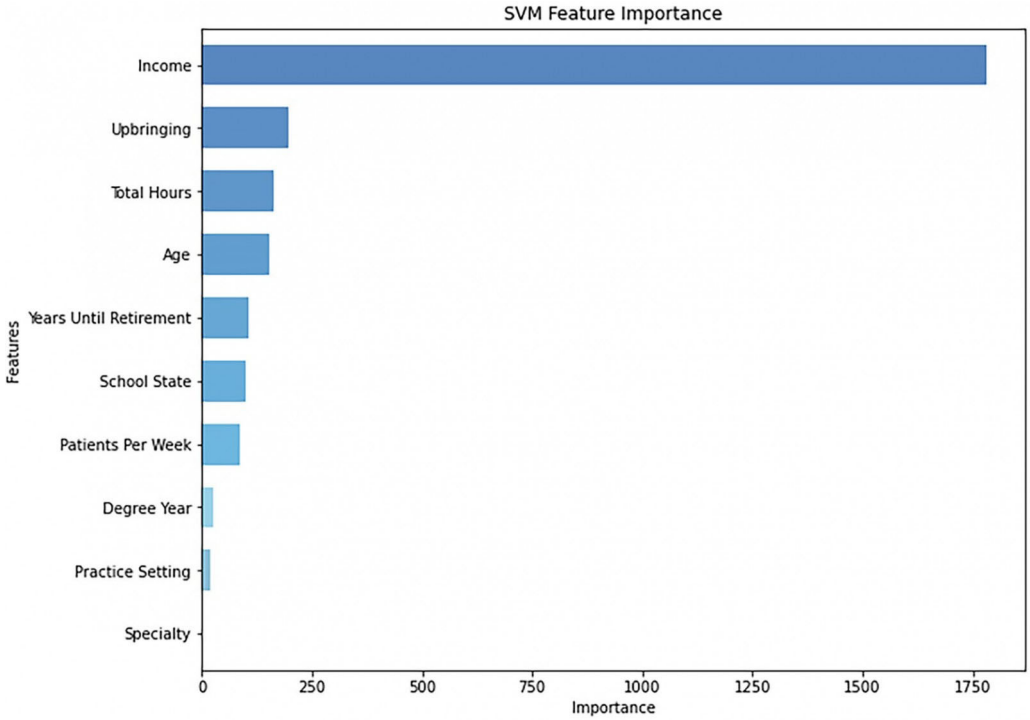

Figure 1: The top 10 most important predictors of practice location from support vector machine.

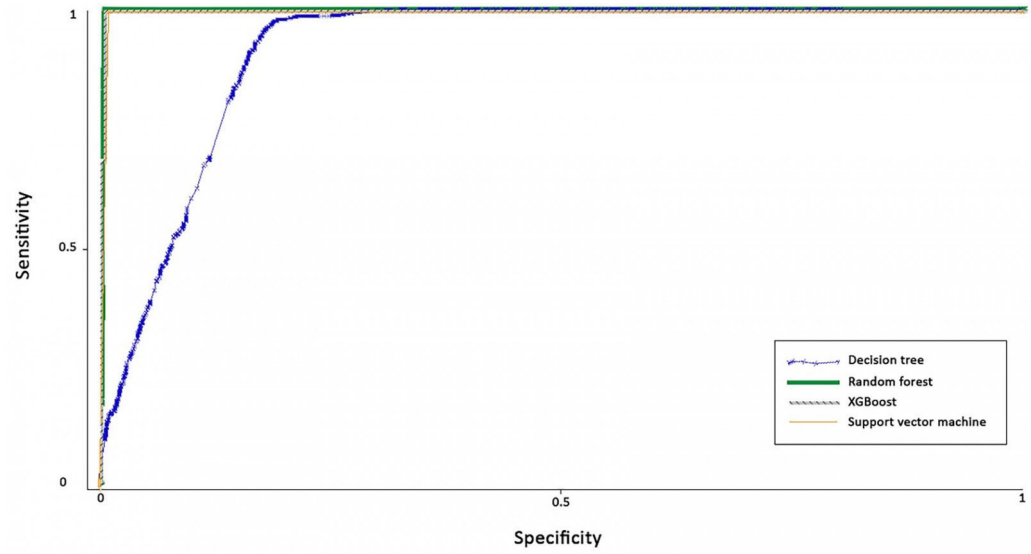

Figure 2: Receiver operating characteristic curves for all models.

Table 2: Descriptive statistics of the top 10 predictor variables of practice location (rural vs non-rural) from support vector machine

Variable	Mean (SD)	Range	*n* (%)
Income (USD)	113,408.00 (97,141.54)	12,131–400,000	
Rural	100,949.22 (90,218.37)	12,131–400,000	
Non-rural	114,802.60 (97,791.79)	12,169–399,997	
Dentist	167,988.40 (91,861.12)	20,430–399,579	
Physician	240,295.09 (98,716.30)	20,910–399,997	
Registered nurse	57,300.57 (26,727.26)	12,131–293,111	
Pharmacist	120,187.60 (37,700.44)	15,000–400,000	
Upbringing			
Rural			705 (63.9)
Non-rural			399 (36.1)
Hours per week	37.51 (14.77)	1–100	
Age (years)	46.59 (12.95)	22–87	
Years until retirement	17.78 (11.11)	1–73	
School state			
Alabama			1 (0.1)
Arizona			15 (1.4)
Arkansas			1 (0.1)
California			25 (2.3)
Colorado			17 (1.6)
Connecticut			2 (0.2)
Florida			7 (0.6)
Georgia			4 (0.4)
Idaho			23 (2.1)
Illinois			11 (1.0)
Indiana			2 (0.2)
International			10 (0.9)
Iowa			11 (1.0)
Kansas			7 (0.6)
Kentucky			11 (1.0)
Louisiana			1 (0.1)
Maryland			3 (0.3)
Massachusetts			3 (0.3)
Michigan			4 (0.4)
Minnesota			4 (0.4)
Missouri			21 (1.9)
Montana			6 (0.5)
Nebraska			35 (3.2)
Nevada			17 (1.6)
New Hampshire			1 (0.1)
New Jersey			2 (0.2)
New Mexico			3 (0.3)
New York			5 (0.5)
North Carolina			2 (0.2)
Ohio			18 (1.6)
Oklahoma			4 (0.4)
Oregon			10 (0.9)
Pennsylvania			10 (0.9)
South Carolina			3 (0.3)
South Dakota			1 (0.1)
Tennessee			6 (0.5)
Texas			15 (1.4)
Utah			736 (67.3)
Vermont			1 (0.1)
Virginia			6 (0.6)
Washington			7 (0.7)
Washington DC			5 (0.5)
Wisconsin			8 (0.7)
Wyoming			8 (0.7)
Patients per week	264.44 (336.44)	0–1800	
Degree year		1960–2017	
Practice setting			
Academic faculty			2 (1.9)
Correctional facility			4 (0.4)
Federally qualified community health center			23 (2.0)
Government agency/armed forces/other federal			3 (0.3)
Home health setting			72 (6.4)
Hospital			538 (48.0)
Insurance claims/benefits			1 (0.1)
Mail order pharmacy			6 (0.5)
Nursing home			49 (4.4)
Occupational health			4 (0.4)
Office/clinic – multi-specialty group			34 (3.0)

Office/clinic – multi-specialty group			64 (5.6)
Office/clinic – single specialty group			95 (8.5)
Office/clinic – solo practice			102 (9.1)
Other			54 (4.8)
Public health			36 (3.2)
Retail pharmacy – chain			24 (2.1)
Retail pharmacy – independent			32 (2.9)
University/college student health facility			15 (1.4)
Volunteer in a free clinic			7 (0.6)
Specialty			
Anesthesiology (general)			6 (0.5)
Cardiology			5 (0.4)
Chronic care			30 (2.7)
Critical care medicine			42 (3.8)
Dermatology			4 (0.4)
Emergency care			85 (7.7)
Gastroenterology			10 (0.9)
General dentistry			91 (8.3)
General pharmacist			82 (7.4)
Hospice and palliative medicine			38 (3.4)
Infectious diseases			12 (1.1)
Labor and delivery			44 (4.0)
Master of Business Administration pharmacist			3 (0.3)
Master of Public Health pharmacist			1 (0.1)
Nephrology			1 (0.1)
Neurology			1 (0.1)
No patient care			52 (4.7)
Occupational health			2 (0.2)
Oncology			4 (0.4)
Ophthalmology			5 (0.5)
Oral and maxillofacial surgery			1 (0.1)
Orthodontics			9 (0.8)
Other dentistry			1 (0.1)
Other specialty			157 (14.2)
Other surgical subspecialties			1 (0.1)
Otolaryngology			3 (0.3)
Pathology (general)			2 (0.2)
Pediatric dentistry			4 (0.4)
Physical medicine and rehabilitation			4 (0.4)
Primary care†			177 (16.1)
Psychiatry			10 (0.9)
Pulmonary disease			1 (0.1)
Radiology (diagnostic)			7 (0.6)
Surgery – general			187 (17.0)
Surgery – orthopedic			10 (0.9)
Urology			9 (0.8)

† Includes general internal medicine, pediatrics and family medicine specialties.
SD, standard deviation.

Table 3: Performance evaluation statistics of the different machine learning methods

Classifier	Accuracy	Precision	Sensitivity	Specificity	AUC	95%CI
Decision tree	0.890	0.834	0.975	0.860	0.914	0.910–0.918
Random forest	0.966	0.937	1.000	0.932	1.000	1.000–1.000
XGBoost	0.966	1.000	0.931	1.000	0.997	0.996–0.998
Support vector machine	0.997	1.000	1.000	0.994	0.997	0.996–0.998

AUC, area under the curve. CI, confidence interval.

Discussion

This study is the first to apply ML techniques to explore factors associated with practising in a rural area. Identifying these factors can facilitate development of effective strategies for recruitment and retention of healthcare professionals into rural settings. By using a healthcare workforce database, ML methods such as DT, RF regression, XGBoost and SVM assessed factors related to rural practice. Among the methods utilized, this study found that SVM worked best in terms of performance in classifying rural practice location. Performance metrics from DT, RF regression and XGBoost also fared well. This experience with a single-state database suggests that ML tools, especially SVM, will be valuable for future research analyzing other state or larger databases that have enough data points to apply ML techniques.

While several studies examine predictors, few assess the relative importance of predictor variables. An important predictor found in this study was upbringing, evidenced by having the second highest importance score (Fig1) and being identified as an important variable linked to rural practice across all the ML methods. This finding is consistent with previous research that also identified rural background as an important predictor of practising in a rural setting[6-11,15-18,24]. By employing four new analytic methods, each

of which reconfirmed that rural upbringing remains linked to choosing a rural practice, this study updates older results and counters concerns[8,25] about the validity of earlier research.

By far, income exhibited the strongest association to practice location. This association aligns with previous studies that found financial factors play a significant role in determining physician practice setting and also for nurses and dentists[54-57]. In the case of physicians, recent salary data indicate that the gap between rural and urban income has narrowed for primary care over the past 5 years[58]. Finding that income connects strongly to rural practice suggests that attractive income packages might help rural communities compete more successfully with urban areas to recruit health professionals and to address shortages. This may be especially important for surgical subspecialties where urban practice remains more lucrative[58]. It also highlights the need for research examining how income influences practice choice and what types of offers are most attractive. Also, while these findings demonstrate linkage to income across four healthcare professions, it may be that flexibility, incentives and other earning features are equally or more important than absolute income. While added income expense may be challenging for rural health professional employers, it might prove to be cost effective if it minimizes turnover and the number and duration of staff vacancies.

Factors with low importance scores in this study also validate earlier research about rural practice choice. Royston et al[12] found that neither gender nor race predicted rural practice, which matched the current findings demonstrated by all four ML methods dropping both features from the final model. The current models also dropped current and total debt as important predictor. Current educational debt being dropped from the model suggests the need for future evaluation on loan repayment incentives. Typically, loan repayment programs offer to pay off a portion of student debt in return for working in an area of high need for a certain period. Both practice setting and specialty had relatively low importance scores compared to the other factors. This finding may be due to rural areas having fewer types of practice settings. Future studies employing ML methods should investigate the association of these factors with rural practice.

Although this study looked at providers in the USA, a disparity of health professionals between urban and rural areas remains a global problem. The finding linking rural upbringing as an important factor for selecting rural practice is similar to international studies examining rural pipelines[59,60]. Although there is substantially more research related to rural pipelines in high-income countries, studies from low- and middle-income countries also demonstrate that students with rural roots are more likely to practise in rural areas[61,62]. Fewer studies explore the impact of income as a factor and most of these focus on physicians and less on other health professions[63]. The current finding identifying income as a strong predictive factor suggests the need for more research about income incentives for allied health professionals and physicians both in high- and lower-income countries.

This study has several limitations. The sample size for each profession differed and was too small to apply ML techniques to each profession separately. Associations identified by this study could be more representative of one healthcare profession over another. However, the models described here provide a method to apply ML techniques to larger databases that have enough data points for separate health professions. Also, examining health professions as a group might be useful for guiding comprehensive strategies to address rural health professional shortages and merits further exploration. Causality from the models is another limitation worth noting. Although the authors identified factors associated with rural practice, causality cannot be inferred. Future studies using ML methods such as causal forest are planned to evaluate causality[64]. A third limitation is that more than half of the original features were dropped due to missing data. A future study using survey questionnaires or databases matched more precisely among the healthcare professions is planned to minimize missing data. Another limitation is how data balance was processed for the training and validation datasets. Although other studies found sampling with replacement to be a reliable method[37-39], over-fitting could still occur when carried out on both datasets. While other techniques to prevent over-fitting, such as balance processing only on the training dataset, exist, sampling with replacement on both the training and validation datasets can reduce bias and model prediction inaccuracy, since it minimizes highly skewed distribution towards healthcare professionals choosing urban practice[40,53,65]. Another limitation was that the sample consisted only of professionals licensed in Utah. Data from other states and internationally will be helpful to confirm model performance and important feature selections. However, Utah represents both large non-rural and rural environments, making it a good state to test ML applications. Also, the times of data collection differed and could affect outcomes. Nonetheless, there is no reason to believe that significant changes occurred over the limited time frame examined. Further research using longitudinal designs is also needed to explore trends. Finally, there are several ways to define 'rural', and this study grouped smaller communities and rural communities into a single designation. In doing this, nuances between a very remote community and a smaller town near a metro area might have been lost.

Conclusion

This study is the first to demonstrate the utility of applying ML methods to identify features linked to rural practice. The study indicates that income is the most important factor associated with rural practice and suggests the need to study what types of income structures might attract more healthcare professionals to rural settings. Rural upbringing emerged as the next most important factor, validating and updating earlier research that identified upbringing as an important factor. Further research applying ML methods to large databases, to explore linkages and to deploy ML algorithms in software applications offer a new tool with the potential to guide and inform strategies that maximize efforts to address rural workforce shortages.

REFERENCES:

1 Johnson GE, Wright FC, Foster K. The impact of rural outreach programs on medical students' future rural intentions and working locations: a systematic review. *BMC Medical Education* 2018; **18(1):** 196. DOI link, PMid:30107795

2 Gondi S, Patel K. Improving rural health: how system-level innovation and policy reform can enhance health outcomes across the United States. *IEEE Pulse* 2016; **7(6):** 8-12. DOI link, PMid:27875111

3 Cecchetti AA. Why introduce machine learning to rural health care? *Marshall Journal of Medicine* 2018; **4(2):** 2. DOI link

4 Redford LJ. Building the rural healthcare workforce: challenges - and strategies - in the current economy. *Generations* 2019; **43(2):** 71-75.

5 Daniels ZM, VanLeit BJ, Skipper BJ, Sanders ML, Rhyne RL. Factors in recruiting and retaining health professionals for rural practice. *Journal of Rural Health* 2007; **23(1):** 62-71. DOI link, PMid:17300480

6 Brooks RG, Walsh M, Mardon RE, Lewis M, Clawson A. The roles of nature and nurture in the recruitment and retention of primary care physicians in rural areas: a review of the literature. *Academic Medicine* 2002; **77(8):** 790-798. DOI link, PMid:12176692

7 Hancock C, Steinbach A, Nesbitt TS, Adler SR, Auerswald CL. Why doctors choose small towns: a developmental model of rural physician recruitment and retention. *Social Science and Medicine* 2009; **69(9):** 1368-1376. DOI link, PMid:19747755

8 Lee DM, Nichols T. Physician recruitment and retention in rural and underserved areas. *International Journal of Health Care Quality Assurance* 2014; **27(7):** 642-652. DOI link, PMid:25252569

9 Rabinowitz HK, Paynter NP. The rural vs urban practice decision. *JAMA* 2002; **287(1):** 113. DOI link

10 Rourke J. How can medical schools contribute to the education, recruitment and retention of rural physicians in their region? *Bulletin of the World Health Organization* 2010; **88:** 395-396. DOI link, PMid:20461207

11 Walker J, DeWitt D, Pallant J, Cunningham C. Rural origin plus a rural clinical school placement is a significant predictor of medical students' intentions to practice rurally: a multi-university study. *Rural and Remote Health* 2012; **12:** 1908. DOI link

12 Royston P, Mathieson K, Leafman J, Sheehan OO. Medical student characteristics predictive of intent for rural practice. *Rural and Remote Health* 2012; **12:** 2107. DOI link, PMid:22873948

13 Rabinowitz HK, Diamond JJ, Markham FW, Paynter NP. Critical factors for designing programs to increase the supply and retention of rural primary care physicians. *JAMA* 2001; **286(9):** 1041-1048. DOI link, PMid:11559288

14 Jones MP, Eley D, Lampe L, Coulston CM, Malhli GS, Wilson I et al. Role of personality in medical students' initial intention to become rural doctors. *Australian Journal of Rural Health* 2013; **21(2):** 80-89. DOI link, PMid:23586569

15 Coyle SB, Narsavage GL. Effects of an interprofessional rural rotation on nursing student interest, perceptions, and intent. *Online Journal of Rural Nursing and Health Care* 2011; **12(1):** 40-48. DOI link

16 McFarland KK, Reinhardt JW, Yaseen M. Rural dentists: does growing up in a small community matter? *Journal of the American Dental Association* 2012; **143(9):** 1013-1019. DOI link, PMid:22942149

17 Scott DM, Neary TJ, Thilliander T, Ueda CT. Factors affecting pharmacists' selection of rural or urban practice sites in Nebraska. *American Journal of Hospital Pharmacy* 1992; **49(8):** 1941-1945. DOI link, PMid:1442837

18 Suphanchaimat R, Cetthakrikul N, Dalliston A, Putthasri W. The impact of rural-exposure strategies on the intention of dental students and dental graduates to practice in rural areas: a systematic review and meta-analysis. *Advances in Medical Education and Practice* 2016; **7:** 623. DOI link, PMid:27822134

19 Lopez N, Sager J, Gonzaga A. Dental and dental therapy students' perspectives on how to build interest in and commitment to rural dentistry. *Journal of Dental Education* 2019; **83(8):** 946-952. DOI link, PMid:31085687

20 McKernan SC, Kuthy RA, Kavand G. General dentist characteristics associated with rural practice location. *Journal of Rural Health* 2013; **29(s1):** s89-s95. DOI link, PMid:23944285

21 Skillman SM, Palazzo L, Keepnews D, Hart LG. Characteristics of registered nurses in rural versus urban areas: implications for strategies to alleviate nursing shortages in the United States. *Journal of Rural Health* 2006; **22(2):** 151-157. DOI link, PMid:16606427

22 Baernholdt M, Mark BA. The nurse work environment, job satisfaction and turnover rates in rural and urban nursing units. *Journal of Nursing Management* 2009; **17(8):** 994-1001. DOI link, PMid:19941573

23 Lea J, Cruickshank M. Supporting new graduate nurses making the transition to rural nursing practice: views from experienced rural nurses. *Journal of Clinical Nursing* 2015; **24(19-20):** 2826-2834. DOI link, PMid:26177875

24 O'Connor SK, Reichard JS, Thrasher KA, Joyner PU. Prospective student pharmacist interest in a rural pharmacy curriculum. *American Journal of Pharmaceutical Education* 2012; **76(6):** 105. DOI link, PMid:22919081

25 Grobler L, Marais BJ, Mabunda S. Interventions for increasing the proportion of health professionals practising in rural and other underserved areas. *Cochrane Database of Systematic Reviews* 2015; **(6):** CD005314. DOI link, PMid:26123126

26 Trickett-Shockey AK, Wilson CS, Lander LR, Barretto GA, Szklarz GD, VanVoorhis GC et al. A study of rural upbringing and education on the intent of health professional students to work in rural settings. *International Journal of Medical Education* 2013; **4:** 18-25. DOI link

27 Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018; **319(13):** 1317-1318. DOI link, PMid:29532063

28 Ruttinger C. *Utah's physician workforce, 2016: a study on the supply and distribution of physicians in Utah.* Salt Lake City, Utah: Utah Medical Education Council, 2016.

29 Caldwell JT, Ford CL, Wallace SP, Wang MC, Takahashi LM. Intersection of living in a rural versus urban area and race/ethnicity in explaining access to health care in the United States. *American Journal of Public Health* 2016; **106(8):** 1463-1469. DOI link, PMid:27310341

30 US Department of Agriculture, Economic Research Service. *2010 Rural-urban commuting area (RUCA) codes* 2014.

31 Njei B, Esserman D, Krishnan S, Ohl M, Tate JP, Hauser RG et al. Regional and rural-urban differences in the use of direct-acting antiviral agents for hepatitis C virus: the veteran birth cohort. *Medical Care* 2019; **57(4):** 279-285. DOI link, PMid:30807449

32 Nagelhout E. *Supply of nurses in Utah: The 2016 Survey of Utah's registered nurses.* 2016. Available: web link (Accessed 7 December 2021).

33 Groesbeck S. *Utah's pharmacist workforce, 2018.* 2018. Available: web link (Accessed 7 December 2021).

34 Christensen J. *Utah's dentist workforce, 2017: a study on the supply and distribution of dentists in Utah.* 2017. Available: web link (Accessed 7 December 2021).

35 Liu Y, Zheng YF. FS_SFS: a novel feature selection method for support vector machines. *Pattern Recognition* 2006; **39(7):** 1333-1345. DOI link

36 Hung M, Voss MW, Rosales MN, Li W, Su W, Xu J et al. Application of machine learning for diagnostic prediction of root caries. *Gerodontology* 2019; **36(4):** 395-404. DOI link, PMid:31274221

37 Viloria A, Lezama OBP, Mercado-Caruzo N. Unbalanced data processing using oversampling: machine learning. *Procedia Computer Science* 2020; **175:** 108-113. DOI link

38 Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making* 2011; **11(1):** 51. DOI link, PMid:21801360

39 Kovács G. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing* 2019; **83:** 105662. DOI link

40 Banerjee P, Dehnbostel FO, Preissner R. Prediction is a balancing act: importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets. *Frontiers in Chemistry* 2018; **6:** 362. DOI link, PMid:30271769

41 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Association for Computing Machinery (Ed.). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17 August 2016, San Francisco.* New York: Association for Computing Machinery, 785-794. DOI link

42 Emanet N, Öz HR, Bayram N, Delen D. A comparative analysis of machine learning methods for classification type decision problems in healthcare. *Decision Analytics* 2014; **1(1):** 6. DOI link

43 Koh HC, Tan G. Data mining applications in healthcare. *Journal of Healthcare Information Management* 2011; **19(2):** 65.

44 Naraei P, Abhari A, Sadeghian A. Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data. *Proceedings of the Future Technologies Conference (FTC), 6-7 December 2016, San Francisco.* New Jersey: IEEE. DOI link

45 Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N et al. Use of machine learning models to predict death after acute myocardial infarction. *JAMA Cardiology* 2021; **6(6):** 633-641. DOI link, PMid:33688915

46 Oztekin A, Delen D, Turkyilmaz A, Zaim S. A machine learning-based usability evaluation method for eLearning systems. *Decision Support Systems* 2013; **56:** 63-73. DOI link

47 Guyon I. *A scaling law for the validation-set training-set size ratio.* Berkeley, California, USA: AT&T Bell Laboratories, 1997.

48 Kannangara M, Dua R, Ahmadi L, Bensebaa F. Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. *Waste Management* 2018; **74:** 3-15. DOI link, PMid:29221873

49 Angra S, Ahuja S. Machine learning and its applications: a review. *Proceedings of the 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)* 23-25 March 2017, Chirala, Andhra Pradesh. New Jersey: IEEE, 2017. DOI link

50 Delen D. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems* 2010; **49(4):** 498-506. DOI link

51 Nevala K. *The machine learning primer: a SAS best practices e-book.* Cary, North Carolina, USA: SAS Institute Inc, 2017.

52 Wehenkel M, Sutera A, Bastin C, Geurts P, Phillips C. Random forests based group importance scores and their statistical interpretation: application for Alzheimer's disease. *Frontiers in Neuroscience* 2018; **12:** 411. DOI link, PMid:30008658

53 Hung M, Hon ES, Ruiz-Negron B, Lauren E, Moffat R, Su W et al. Exploring the intersection between social determinants of health and unmet dental care needs using deep learning. *International Journal of Environmental Research and Public Health* 2020; **17(19):** 7286. DOI link, PMid:33036152

54 Goodfellow A, Ulloa JG, Dowling PT, Talamantes E, Chheda S, Bone C et al. Predictors of primary care physician practice location in underserved urban and rural areas in the United States: a systematic literature review. *Academic Medicine: Journal of the Association of American Medical Colleges* 2016; **91(9):** 1313. DOI link, PMid:27119328

55 Mbemba G, Gagnon M-P, Paré G, Côté J. Interventions for supporting nurse retention in rural and remote areas: an umbrella review. *Human Resources for Health* 2013; **11(1):** 44. DOI link, PMid:24025429

56 Skillman SM, Hager LJ, Frogner BK. *Incentives for nurse practitioners and registered nurses to work in rural and safety net settings.* University of Washington Centre for Health Workforce Studies Rapid Turnaround Brief. November 2015.

57 Werts MA, Amah G, Mertz E. *How evidence-based is US dental workforce policy for rural communities?* Rensselaer, New York, USA: Oral Health Workforce Research Center, Center for Health Workforce Studies, 2020; 52.

58 Darves B. Demystifying urban versus rural physician compensation. *New England Journal of Medicine CareerCenter* 4 March 2019. Available: web link (Accessed 16 December 2020).

59 Dunbabin JS, Levitt L. Rural origin and rural medical exposure: their impact on the rural and remote medical workforce in Australia. *Rural and Remote Health* 2003; **3(1):** 212. DOI link, PMid:15877502

60 Silvestri DM, Blevins M, Afzal AR, Andrews B, Derbew M, Kaur S et al. Medical and nursing students' intentions to work abroad or in rural areas: a cross-sectional survey in Asia and Africa. *Bulletin of the World Health Organization* 2014; **92:** 750-759. DOI link, PMid:25378729

61 Zimmerman M, Shakya R, Pokhrel BM, Eyal N, Rijal BP, Shrestha RN et al. Medical students' characteristics as predictors of career practice location: retrospective cohort study tracking graduates of Nepal's first medical college. *BMJ* 2012; **345:** e4826. DOI link, PMid:22893566

62 Henry JA, Edwards BJ, Crotty B. Why do medical graduates choose rural careers? *Rural and Remote Health* 2009; **9(1083):** 1-13. DOI link, PMid:19257797

63 Kumar S, Tian EJ, May E, Crouch R, McCulloch M. 'You get exposed to a wider range of things and it can be challenging but very exciting at the same time': enablers of and barriers to transition to rural practice by allied health professionals in Australia. *BMC Health Services Research* 2020; **20(1):** 1-14. DOI link, PMid:31888624

64 Athey S, Tibshirani J, Wager S. Generalized random forests. *The Annals of Statistics* 2019; **47(2):** 1148-1178. DOI link

65 Hung M, Li W, Hon ES, Su S, Su W, He Y et al. Prediction of 30-day hospital readmissions for all-cause dental conditions using machine learning. *Risk Management and Healthcare Policy* 2020; **13:** 2047-2056. DOI link, PMid:33116985

APPENDIX I:

Appendix I: The top 28 features from random forest feature selection

Variable	Description
Provide services in Utah†	Practising in or out of Utah
Profession†	Type of healthcare profession
Gender†	Gender (male or female)
Age	Age in years
Ethnicity†	Hispanic or non-Hispanic ethnicity
Race†	Respondent's race
Upbringing	Area (rural, urban, suburban) respondent spent the majority of their upbringing
Utah upbringing†	If the respondent spent the majority of their upbringing outside or inside Utah
State upbringing†	The state the respondent spent most of their upbringing in
Degree type†	Type of health profession degree
School state	The state the health profession degree was conferred
Institution†	The institution the health profession degree was conferred
Degree year	The year the health profession degree was conferred
Current debt†	The current debt of the respondent
Total debt†	The overall debt the respondent accumulated
Income	Annual salary
Primary specialty	The primary specialty the respondent practises
Primary hours†	Number of hours the respondent spends in their primary practice
Primary direct patient care hours†	Number of hours the respondent spends on direct patient care
Average hours†	Average hours the respondent spends a week
Total hours	Total number of hours the respondent spends in all practice settings
Primary setting	The primary work setting of the respondent
Switched employers†	Whether the respondent switched employers in the last five years
Retirement age†	The age in years the respondent plans to retire
Years until retirement	The estimated years the respondent has until retirement
Patients per week	The number of patients the respondent sees per week
New patient wait times†	The wait time (in hours) new patients have to wait for an appointment
Established patient wait times†	The wait time (in hours) established patients have to wait for an appointment

† Dropped variables not included in the final model.

State upbringing†	The state the respondent spent most of their upbringing in
Degree type†	Type of health profession degree
School state	The state the health profession degree was conferred
Institution†	The institution the health profession degree was conferred
Degree year	The year the health profession degree was conferred
Current debt†	The current debt of the respondent
Total debt†	The overall debt the respondent accumulated
Income	Annual salary
Primary specialty	The primary specialty the respondent practises
Primary hours†	Number of hours the respondent spends in their primary practice
Primary direct patient care hours†	Number of hours the respondent spends on direct patient care
Average hours†	Average hours the respondent spends a week
Total hours	Total number of hours the respondent spends in all practice settings
Primary setting	The primary work setting of the respondent
Switched employers†	Whether the respondent switched employers in the last five years
Retirement age†	The age in years the respondent plans to retire
Years until retirement	The estimated years the respondent has until retirement
Patients per week	The number of patients the respondent sees per week
New patient wait times†	The wait time (in hours) new patients have to wait for an appointment
Established patient wait times†	The wait time (in hours) established patients have to wait for an appointment

† Dropped variables not included in the final model.