RESEARCH

Computer-Assisted Decision Support for Student Admissions Based on Their Predicted Academic Performance

Eugene Muratov, PhD,[a] Margaret Lewis, MEd,[a] Denis Fourches, PhD,[b] Alexander Tropsha, PhD,[a] Wendy C. Cox, PharmD[a]

[a] UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina
[b] Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina

Objective. To develop predictive computational models forecasting the academic performance of students in the didactic-rich portion of a doctor of pharmacy (PharmD) curriculum as admission-assisting tools.

Methods. All PharmD candidates over three admission cycles were divided into two groups: those who completed the PharmD program with a GPA $\geq$ 3; and the remaining candidates. Random Forest machine learning technique was used to develop a binary classification model based on 11 pre-admission parameters.

Results. Robust and externally predictive models were developed that had particularly high overall accuracy of 77% for candidates with high or low academic performance. These multivariate models were highly accurate in predicting these groups to those obtained using undergraduate GPA and composite PCAT scores only.

Conclusion. The models developed in this study can be used to improve the admission process as preliminary filters and thus quickly identify candidates who are likely to be successful in the PharmD curriculum.

Keywords: admissions, academic performance, Pharmacy College Admission Test, computer-mediated communication, evaluation methodologies

INTRODUCTION

Admissions committees of health sciences schools spend a significant amount of time and effort screening applications to assess candidates' readiness for a professional program. Competition for admission is very high with more applicants than available slots, so both rigor and efficiency are valued in the admissions process. Often admissions committees attempt to predict an applicant's academic success as a means of assisting the admission decision by selecting candidates most likely to perform well in the didactic portion of the professional curriculum.[1-4] Such models typically rely on linear or multiple linear regression approaches based on a few parameters such as applicants' cumulative grade point average (GPA), perceived rigor of prior university attended, scientific rigor of classes taken, and standardized test scores. However, hidden patterns or relationships between variables cannot be evaluated with these models and are often over simplified.[5]

Non-linear multivariate models are widely used to forecast various types of phenomena. These models capture hidden relationships between variables (also called attributes or descriptors) and an endpoint of interest.[6] Such models are common in many areas of pharmaceutical research, such as computer-aided drug discovery.[7] There have been some attempts to use more complex computational techniques to build models that assist in making admission decisions, predict the academic success of students, or predict admission yields with varying results.[5,8,9] For instance, students' Medical College Admission Test (MCAT) scores and undergraduate GPA were considered indicators of success in the classroom;[2,10] however, other variables and combinations of these parameters were found to contribute to a student's success in medical school didactic curricula.[11-13] Previous studies have also explored the association of admission parameters with academic success in a pharmacy school, as measured by pharmacy GPA, using factors such as performance on the Pharmacy College Admission

Corresponding Author: Wendy C. Cox, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, CB 7566, 100J Beard Hall, Chapel Hill, NC 27599-7566. Tel: 919-966-9990. E-mail: wendy_cox@unc.edu

Test (PCAT) and undergraduate GPA.[3,14-16] Although we know PCAT scores and undergraduate GPA can predict success in the didactic-rich portion of pharmacy programs, other parameters often factor into admission decisions. Some of these factors are the undergraduate university attended, undergraduate college major, presence of a prior 4-year degree, and various non-cognitive attributes.

Our school is undergoing a curriculum transformation with plans to implement several new teaching and learning strategies, shorten foundational coursework, and enable earlier immersion into patient care roles in order to meet the health care needs of the future.[17] As part of the transformation, a new admissions process has been developed with the goal of selecting students who will likely be successful in the new curriculum and in pharmacy practice. It is understood that such students must possess certain non-cognitive attributes, including resiliency, adaptability, and empathy. However, at the initial stage of the admissions process, we still need to efficiently select candidates who are expected to be well prepared for the curriculum, so more time can be spent evaluating the non-cognitive attributes of candidates.

The objectives of this study were to enhance the admissions process by using computational models capable of accurately forecasting the academic performance of students in the didactic-rich portion of the doctor of pharmacy (PharmD) curriculum. We tried to develop classification models using both historical admissions and academic performance data for cohorts of current students in our PharmD program. Our models can accurately predict the academic performances of all PharmD candidates in two categories, ie, highest performing and lowest performing students. As a proof of their pragmatic value, these models were piloted in the UNC Eshelman School of Pharmacy as decision-support tools in the 2014-2015 admission cycle.

METHODS

Each year, approximately 170 PharmD candidates are selected for admission at the UNC Eshelman School of Pharmacy from a pool of over 650 applicants. For each applicant, we evaluated several academic parameters of their prior education, including GPA, PCAT composite score, and PCAT subscores as well as undergraduate major, presence of a four-year degree, rigor of the undergraduate university attended, and performance in select coursework. How these parameters contribute to a candidate's success in the didactic-rich portion of the curriculum is much more difficult to assess. If the candidate met a minimum academic threshold where we felt confident that they can be successful academically, their application

package (personal essay, extracurricular and work experience, honors and awards, and letters of recommendation) was sent to a formal application review where their selected professional and non-cognitive attributes, often called "soft skills," were evaluated. Candidates who achieved a minimum score in the first and second stage of the admissions process were invited for an on-campus multiple mini-interview (MMI) where their non-cognitive attributes were assessed. Following the MMI, the admissions committee makes the final admission decision.

All candidates who applied during the 2007-2008 (graduating class of 2012), 2008-2009 (class of 2013), and 2009-2010 (class of 2014) admission cycles were included in the study. The model was built using 1,389 applicant records (742 from the 2007-2008 admission cycle and 647 from the 2008-2009 admission cycle). Admitted candidates who later earned a PharmD GPA ≥ 3 were assigned to category 1 and admitted candidates who earned a PharmD GPA < 3, and those who were denied admission were assigned to category 0. In total there were 253 candidates in category 1 and 1,136 candidates in category 0. Since we have united the data from both 2007-2008 and 2008-2009 admission cycles for developing our models, we refer to them throughout this article as the 2007-2009 cohort. There were 713 candidates from the 2009-2010 admission cycle (class of 2014) who were used for additional external validation of developed models.

Eleven pre-admissions parameters were considered in this study: undergraduate GPA (uGPA), combination of undergraduate and graduate GPA (ugGPA), presence of a four-year degree, undergraduate major, selectivity index of the undergraduate school, PCAT composite score, and PCAT percentile subscores (biology, chemistry, quantitative, reading comprehension, and verbal). The highest PCAT scores across multiple test dates were used for both the composite score and subscores. The uGPA was calculated by examining all undergraduate level classes completed, whereas the ugGPA was calculated by examining all undergraduate and graduate level coursework completed. The presence of a four-year prior degree was examined and the undergraduate majors were grouped into five categories: chemistry/biochemistry, biology/microbiology, other science, non-science, and pre-pharmacy/no major.

We also considered a metric characterizing the rigor of training in the college attended by applicants. It is difficult to capture quantitatively; however, the selectivity index (a measure of how difficult it is for students to gain admission) is often used. The selectivity index of an institution was determined by calculating the number of

admitted candidates divided by the total applicant pool. Institutional admissions data were taken from the Integrated Postsecondary Education Data System (IPEDS) Data Center.[18]

We used a Random Forest (RF) algorithm[19] that we implemented earlier.[20] RF is an ensemble of decision trees whose outputs are aggregated to obtain one final prediction (here, PharmD GPA) by majority voting.[19] Each tree is grown as follows: (i) a bootstrap sample forming the training set for the current tree is produced from the whole training set of N cases. Cases that are not in the current tree training set are placed in an out-of-bag set (~N/3 cases). (ii) The best split among the M randomly selected parameters from the initial set is chosen in each node by the CART algorithm.[21] The value of M is just one tuning parameter to which RF models are sensitive. (iii) Each tree is grown to the largest possible extent without any pruning. The models were selected as part of the ensemble according to their performance on an out-of-bag set.[22]

We rigorously validated our models using the 5-fold external cross-validation procedure: the full set of applicants with assigned category (0 or 1) was randomly divided into five subsets of equal size; then one of these subsets (20% of all cases) was set aside as an external validation set and the remaining four sets together formed the modeling set (80% of the full set). This procedure was repeated five times allowing each of the five subsets to be used as an external validation set. Models were built using the modeling set only, and it is important to emphasize that the external set cases were never used either to build and/or select the models. Each modeling set was divided into many internal training and test sets; then models were built using cases of each training set and applied to the test set.

The following statistical characteristics were used for assessing the model performance: sensitivity (SE), which is the ratio of the number of applicants correctly predicted to earn the higher PharmD GPAs (ie, in category 1 as defined in the previous section) to the total number of students with higher PharmD GPAs; specificity (SP), which is the ratio of the number of applicants correctly predicted either to earn the lower PharmD GPAs or not admitted (ie, in category 0) to the total number of lower PharmD GPAs; Correct Classification Rate (CCR) is the average of SP and SE; Positive Prediction Value (PPV) is the probability of correct prediction of students with the higher PharmD GPAs, which is the ratio of number of applicants earning the higher PharmD GPAs divided by the sum of the number of applicants correctly predicted to earn the higher PharmD GPAs and the number of applicants wrongly

predicted to earn the higher PharmD GPAs; Negative Prediction Value (NPV) is the probability of correct prediction of lower PharmD GPAs or not admitted, which is the ratio of the number of applicants earning the PharmD GPAs or not admitted divided by the sum of the number of applicants correctly predicted to earn the lower PharmD GPAs or not admitted and the number of applicants wrongly predicted to earn the lower PharmD GPAs or not admitted.

This study was submitted and considered exempt from further review by the Institutional Review Board of the University of North Carolina at Chapel Hill.

RESULTS

The overall study design is shown in Figure 1. In brief, we (i) collected and (ii) rigorously curated the data; (ii) divided the applicants into in-state residents vs. non-residents and developed separate models for these two groups; (iii) interpreted the models to estimate the importance of applicants' characteristics (descriptors); and (iv) used developed models for the prescreening of new applicants to select those academically prepared for a formal and in-depth application review of non-cognitive and professional attributes. These steps are described in more detail below.

As we have shown previously, data curation is an obligatory part of any modeling.[23,24] Thus, before building the models, we carefully evaluated the input data for completeness and consistency. As a result, 32 students were excluded from the 2009-2010 cohort because of missing descriptor values. In addition, we found that 84 students from the 2007-2009 cohort had erroneous college selectivity index values. Most of these students completed their prior degree outside of the United States and their institutions were not assigned the selectivity index; however, in our input data, their SI values were artificially recorded as zero. Since missing or erroneous descriptor values are unacceptable for modeling, all of these records were removed.

We also took into account the distribution of applicants based on their residency status. As a state-supported school, we must accept significantly more resident or in-state candidates (60%-80%) than non-resident or out-of-state ones (20%-40%). Thus, there is more competition for the admission within, as opposed to between, these two groups of applicants. Therefore, we decided to develop two separate models: model 1 based on combined 2007-2008 and 2008-2009 admission data (2007-2009 cohort) for resident candidates; and model 2 based on 2007-2008 and 2008-2009 (2007-2009 cohort) admission data for non-resident candidates. The final number of records used to build and validate models 1 and 2 (after
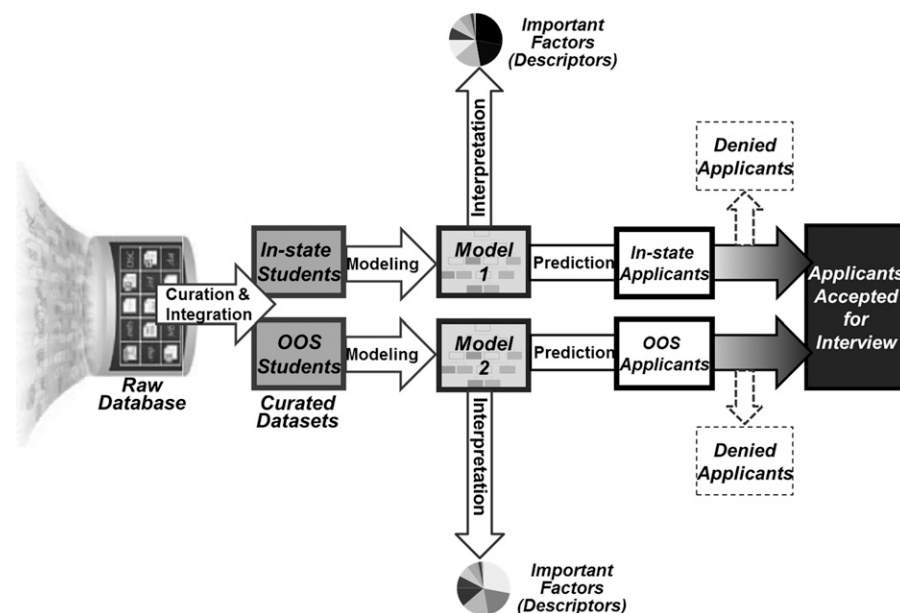
Figure 1. Study Design of Computational Model Development and Use.

excluding those for candidates with missing records or erroneous selectivity index data) are listed in Table 1. Both models 1 and 2 were additionally validated using the curated 2009-2010 admissions data.

The data for the 2007-2009 cohort included 406 resident candidates (191 in category 1 with PharmD GPAs $\geq$ 3.0 and 215 in category 0 with PharmD GPAs $<$ 3.0 or denied admission). The model validation dataset for the 2009-2010 cohort involved 162 candidates (90 with PharmD GPAs $\geq$ 3.0 and 72 with PharmD GPAs $<$ 3.0 or denied admission). Statistical characteristics of the developed models estimated using both 5-fold external CV and the external set (see Methods) are listed in Table 2. We succeeded in developing a robust and externally predictive model for in-state candidates. In general, the statistics for the modeling set (2007-2009 cohort) were somewhat better than that for the external set of 2009-2010 candidates; however, the values for all statistical characteristics were nearly the same: ranges of CCR, sensitivity, and specificity were 74%-77% vs. 63%-70%, respectively.

The admissions process for non-resident candidates is very competitive due to a large number of applicants for a small number of slots, and thus the data were highly unbalanced for the 2007-2009 cohort (54 admitted vs. 844 denied admission). A similar situation was observed for the 2009-2010 cohort of non-resident candidates (28 admitted vs. 491 denied). Initially, we did not succeed in developing a significant model for this unbalanced dataset (results not shown). Thus, we balanced the dataset using all 54 candidates admitted and 57 randomly chosen

candidates who were denied admission. The remaining 787 candidates denied admission were used as an additional external validation set. All 2009-2010 admission data were used for external validation as well (see Table 1

Table 1. Number of Applicants Used in the Modeling

	High	Low
Models developed using all 11 parameters.		
Model 1 2007-2009	191	215
Model 1 2009-2010	90	72
Model 2 2007-2009	54	57
Model 2 2007-2009 excluded set[a]	0	787
Model 2 2009-2010	28	491
Model 1 2007-2009 extreme cases[b]	27	33
Model 1 2009-2010 extreme cases[b]	17	13
Model 2 2007-2009 extreme cases[b]	13	0
Model 2 2009-2010 extreme cases[b]	8	4
Models developed using only undergraduate and graduate GPA and composite PCAT.		
Model 1a 2007-2009	191	215
Model 1a 2009-2010	90	72
Model 2a 2007-2009	54	57
Model 2a 2007-2009 excluded set[a]	0	787
Model 2a 2009-2010	28	491
Model 1a 2007-2009 extreme cases[b]	27	33
Model 1a 2009-2010 extreme cases[b]	17	13
Model 2a 2007-2009 extreme cases[b]	13	0
Model 2a 2009-2010 extreme cases[b]	8	4

[a]Model prediction of 787 denied candidates
[b]Extreme cases are the following: PharmD GPA $\geq$ 3.8 for high performance students and PharmD GPA $<$ 3 for low performance students

Table 2. Statistical Characteristics of Developed Models

Models	Statistical parameters, %				
	PPV	NPV	SE	SP	CCR
Models developed using all 11 parameters					
Model 1 2007-2009	73	79	77	74	76
Model 1 2009-2010	70	63	70	63	66
Model 2 2007-2009	78	85	85	77	81
Model 2 2007-2009 excluded set [a]	N/A	N/A	N/A	75	N/A
Model 2 2009-2010	17	98	75	79	77
Model 1 2007-2009 extreme cases [b]	67	88	89	64	76
Model 1 2009-2010 extreme cases [b]	79	82	88	69	79
Model 2 2007-2009 extreme cases [b]	100	N/A	85	N/A	85
Model 2 2009-2010 extreme cases [b]	86	60	75	75	75
Models developed using only undergraduate and graduate GPA and composite PCAT					
Model 1a 2007-2009	67	73	71	69	70
Model 1a 2009-2010	64	52	56	61	58
Model 2a 2007-2009	76	86	87	74	80
Model 2a 2007-2009 excluded set [a]	N/A	N/A	N/A	77	N/A
Model 2a 2009-2010	17	98	68	82	75
Model 1a 2007-2009 extreme cases [b]	58	77	81	52	66
Model 1a 2009-2010 extreme cases [b]	69	53	53	69	61
Model 2a 2007-2009 extreme cases [b]	100	N/A	76	N/A	76
Model 2a 2009-2010 extreme cases [b]	83	50	63	65	69

PPV = Positive prediction value (probability of correct prediction of students with higher PharmD GPAs)
NPV = Negative prediction value (probability of correct prediction of students with lower PharmD GPAs or not admitted)
SE = Sensitivity (ratio of the number of applicants correctly predicted to earn a higher PharmD GPA)
SP = Specificity (ratio of the number of applicants correctly predicted to earn a lower PharmD GPAs or not admitted)
CCR = Correct classification rate (average of SP and SE)
[a]Model prediction of 787 candidates denied admission
[b]Extreme cases: PharmD GPA $\geq$ 3.8 for high performance students and PharmD GPA $<$ 3 for low performance students

for details). Statistical characteristics of the developed models are shown in Table 2. We succeeded in developing robust and predictive models for the non-resident candidates. The prediction accuracy for the 787 candidates denied admission was as high as 75%. Similar to model 1, the statistics for the modeling set (2007-2009 cohort) were only slightly higher than for the external set of 2009-2010 cohort (ranges of CCR, sensitivity, and specificity were 77%-85% vs. 75%-79%, respectively).

Our models could be particularly useful for detecting students belonging to extreme groups. The first group represented the students with PharmD GPA $\geq$ 3.8 (in total, 44 records for residents and 21 records for non-residents) and the second group with PharmD GPA $<$ 3 (in total, 46 records for residents and 4 records for non-residents). More details about the population of marginal groups are given in Table 1 for models marked as "extreme cases." We applied models 1 and 2 to predict the performances of resident and non-residents in both extreme groups with the following results: 24 out of 27 residents with PharmD GPA $\geq$ 3.8 from the 2007-2009 cohort and 15 out of 17 from the 2009-2010 cohort were

predicted correctly; 21 out of 33 residents with PharmD GPA $<$ 3 from the 2007-2009 cohort and 9 out of 13 from the 2009-2010 cohort were predicted correctly; 11 out of 13 non-residents with PharmD GPA $\geq$ 3.8 from the 2007-2009 cohort and 6 out of 8 from the 2009-2010 cohort were predicted correctly; 21 out of 33 non-residents with PharmD GPA $<$ 3 from the 2007-2009 cohort and 3 out of 4 non-residents with PharmD GPA $<$ 3 from the 2009-2010 cohort were predicted correctly (2007-2009 cohort does not include non-residents with PharmD GPA $<$ 3). Statistical characteristics for the prediction of the extreme groups can be found in Table 2 for models marked as "extreme cases."

In general, in-state residents with a high GPA were better predicted compared to in-state residents with a low PharmD GPA. The quality of predictions for non-residents was comparable to that for residents. At the same time, the small number of non-residents with low academic performance (only four in three years) does not allow us to make any robust conclusions about the prediction reliability for students with a low PharmD GPA. Overall, in the entire analyzed pool of candidates, there

were 115 students belonging to extreme groups and 89 of them were predicted correctly, ie, we achieved an external accuracy of 77%.

We aimed to address the question if the use of 11 descriptors (see Methods) was justified or ugGPA and composite PCAT, the most obvious and commonly used *a priori* predictors of student performance used in some other studies [1,25,26] could predict PharmD GPA with similar accuracy. Our analysis of relative significance of all descriptors used in modeling (Table 3) suggested that ugGPA and composite PCAT were indeed the most important characteristics. However, the correlation between ugGPA alone and PharmD GPA was weak (r=0.47, Figure 2). To further address this question, we have built models 1a and 2a, which were analogous to models 1 and 2, respectively, using only composite PCAT and ugGPA (Tables 1 and 2). The comparison of statistical characteristics indicated that our model built with all 11 descriptors outperformed the two-descriptor models for resident candidates (CCR of 66%-76% vs. 58%-70% for model 1 vs. model 1a, respectively). Predictive performances of models 2 and 2a developed for non-resident candidates were very similar with 11-descriptor model being slightly more accurate: CCR of 77%-81% vs. 75%-80% for model 2 vs. model 2a, respectively. At first sight, 11 descriptor models, especially for out-of-state applicants, provide only marginal advantage over two descriptor models. However, prediction of the extreme cases shows their clear and substantial advantage. As seen from Table 2, in all cases, performance of two descriptor models was significantly lower than the performance of their 11 descriptor analogs. In-state residents with a high PharmD GPA from the 2007-2009 cohort and residents with a low PharmD GPA from the 2009-2010 cohort were predicted especially poorly by two descriptor models with an accuracy close to 50%. These results once again in-dicate the benefits of using all 11 descriptors. The significance of the findings of our study are summarized in Table 4.

DISCUSSION

Several models have been previously developed to assist admission officers in decision making based on predicted academic success.[1,25-27] However, to the best of our knowledge this is the first report where multivariate statistical models using most available academic admissions data were developed as an admission decision-support tool. We consider this approach novel due to the concurrent use of multiple student admission parameters to forecast academic performance, as opposed to using only a few parameters. It is difficult to simultaneously consider 11 parameters, their relationship to one another, and the proper weight of each parameter. This is an objective way admission officers can consolidate diverse data and objectively evaluate the academic readiness of applications. Using admitted student data, we have developed a series of computational models that, with high accuracy, predicted academic performance of students in the PharmD curriculum at UNC, especially for best performing students earning GPAs $\geq$ 3.8 or for those with relatively low GPAs < 3.0. Notably, multivariate models using 11 pre-admission student descriptors performed significantly better than those built with only two most obvious applicant characteristics such as ugGPA and composite PCAT.

Previous tools designed to assist with the admissions process have relied typically on linear regression models based on standardized test scores and entering GPAs.[1,25–27] Factors such as rigor of prior university attended, college major, presence of a four year degree, and grades in select coursework contribute to a candidate's ability to be successful academically, but reports of models using all

Table 3. Relative Descriptor Importance

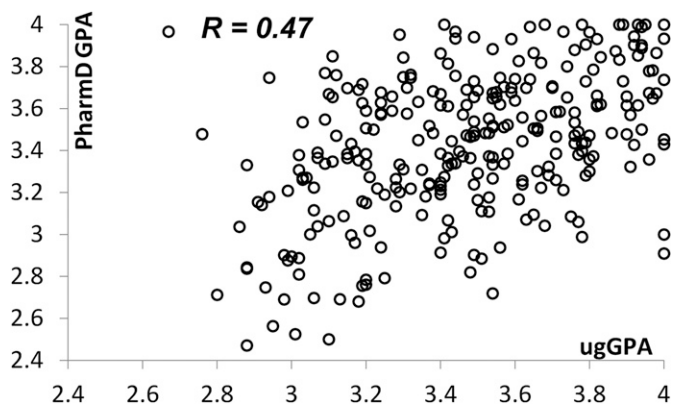| Variable name	Model 1[a]	Model 2[a]
Biology PCAT	3	5
Chemistry PCAT	4	5
Composite PCAT	27	44
Major	0	0
Previous Degree	1	0
Quantitative PCAT	8	14
Reading Comprehension PCAT	2	7
Selectivity Index	6	0
Undergraduate & Graduate GPA	21	11
Undergraduate GPA	23	13
Verbal PCAT	5	1

[a]Percentage

Figure 2. Weak Correlation (R) Between PharmD GPA (y axis) and Undergraduate and Graduate (ug) GPA (x axis).

Table 4. Significant Findings of the Study

Finding	Significance
Eleven descriptors provide more accurate predictions than undergraduate and graduate GPA and composite PCAT	This resulted in the decreased risk of error of prediction, ie, decreased chance of rejection of applicants who are academically prepared or moving applicants with low academic performance to the next step in the admission process.
Separate models were built for in-state residents and non-residents	This resulted in the increased accuracy of prediction for both in-state residents and non-residents.
Important descriptors different for residents and non-residents were identified	This allowed us to focus on the most important characteristics for academic performance during selection process and the interview.
Predictive models were developed and used for virtual screening of new applicants	Use of the models reduced the time spent evaluating academics by approximately two-thirds.
	As a result of using the models, we had more confidence in identifying students predicted to be high or low performers in the curriculum and our academic evaluation was more efficient.
Overall study impact	Admission officers can develop a specific model for their school using the approach described in this study with more accuracy than using GPA and PCAT scores only. Admissions officers can easily and quickly input the admission parameters for new applicants into the software and receive an output of high, low, or undetermined predicted performance. These models would allow admissions offices to filter out applicants who are not academically prepared for their program, leaving more time to review the non-cognitive or professional attributes of academically prepared applicants in more depth.

these data in combination to predict student academic performance have been uncommon. As in the previous studies,[3,14–16,25–27] the undergraduate GPA and composite PCAT scores were found to be the most impactful (Table 3). However, the consensus use of all 11 descriptors in our models resulted in a significantly more accurate prediction of PharmD GPA category, particularly for extreme cohorts of highest- and lowest-performing students than the traditional filter based only on composite PCAT scores and undergraduate GPA (Table 2). Surprisingly, college major was not found to influence the model (Table 3) and thus it could be eliminated from our model. This may be due to the requirement that applicants complete specific math and science pre-requisite courses prior to entering our program. Other programs that do not require pre-requisite courses may yield different results.

As a public, state-supported school, we accept more in-state or resident candidates (60%-80% per entering class) whereas the majority (60%-70%) of applicants are non-residents. For this reason, we have developed and validated two separate models: one to predict the performance of residents and another to predict the performance of non-residents. Both models were built using the same descriptors and protocol. The major difference between them are the datasets used for modeling, ie, in-state residents vs. non-residents. This resulted in the different statistical characteristics of the models and in different relative importance of

descriptors. These models were piloted during the 2014-2015 admission cycle in our school with positive results. Specifically, the use of these models resulted in less time spent evaluating academic preparedness and a reduction in non-academically qualified applicants sent to stage 2 of the admissions process (formal application review), which is very time-intensive and is completed by trained application readers. However, we did have to individually review those applicants that the models were unable to evaluate with acceptable accuracy, ie, those that were not predicted to belong to either of the two extreme cohorts with GPA ≥ 3.8 or GPA < 3.0. Although the exact number of hours saved was not recorded, it is estimated that this process reduced the time spent evaluating academic credentials of applicants by two-thirds, with the majority of the time spent evaluating the applicants that the models were not able to predict with the acceptable accuracy. As a result of using the models, we had more confidence in students predicted to be high or low performers in the curriculum and our academic evaluation was more efficient.

We expect the approach developed in this study can be easily applied to other pharmacy and health profession schools and even to undergraduate or graduate admissions because most of the descriptors used in our study are commonly available for most applicants. The implementation of these models in any school or

college requires a dedicated person to collect and clean admission and grade data, and a skilled statistician to perform the modeling. Once the specific model for the school is developed using the approach described in this study, admissions officers can easily and quickly input the admission parameters for new applicants into the software and receive an output of high, low, or undetermined predicted performance. This model would allow admissions offices to filter out applicants who are not academically prepared for their program, leaving more time to review the non-cognitive or professional attributes of academically prepared applicants in more depth.

Our models do have some limitations. First, all 11 descriptors must be provided for every applicant as missing data will result in inaccurate results. Second, the reliable prediction of actual PharmD GPA values is another unsolved challenge due to their very narrow range (2.8-4) with the majority of data clustered within the 3-3.85 interval. Enriching our database by students graduating in 2015 and subsequent years and addition of new descriptors characterizing the applicants could help solve this problem. Future studies will also examine the validity of this approach for exploring non-cognitive attributes of student preparation and training. We also envision using both pre-admission data and performance in the professional curriculum for forecasting professional success of students following their graduation.

CONCLUSION

We have developed a series of computational models that employ admissions data to forecast students' performance (assessed by PharmD GPA) in the didactic-rich portion of a PharmD program. Models developed in this study were used by the admissions office as preliminary filters to quickly select subset(s) of candidates likely to be successful in the didactic-rich portion of the PharmD curriculum. As a result, more time could be spent evaluating these candidates' non-cognitive attributes and their overall fit for the program.

The models developed with all 11 descriptors afford significantly higher prediction accuracy than simple estimates based on the ugGPA and composite PCAT, especially in predicting applicants with the highest (≥ 3.8) and the lowest (< 3) PharmD GPAs.

ACKNOWLEDGMENTS

REFERENCES

1. Kuncel NR, Credé M, Thomas LL, Klieger DM, Seiler SN, Woo SE. A meta-analysis of the validity of the Pharmacy College Admission Test (PCAT) and grade predictors of pharmacy student performance. *Am J Pharm Educ*. 2005;69(3):Article 51.
2. Dunleavy DM, Kroopnick MH, Dowd KW, Searcy CA, Zhao X. The predictive validity of the MCAT exam in relation to academic performance through medical school: a national cohort study of 2001-2004 matriculants. *Acad Med*. 2013;88(5):666-671.
3. Allen DD, Bond CA. Prepharmacy predictors of success in pharmacy school: grade point averages, Pharmacy College Admissions Test, communication abilities, and critical thinking skills. *Pharmacotherapy*. 2001;21(7):842-849.
4. Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. *Adv Health Sci Educ Theory Pract*. 2009;14(5):759-775.
5. Chang L. Applying data mining to predict college admissions yield: a case study. *New Dir Institutional Res.* 2006;2006(131):53-68.
6. Randić M. Generalized molecular descriptors. *J Math Chem.* 1991;7(1):155-168.
7. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? *J Med Chem.* 2014;57 (12):4977-5010.
8. Gorr WL, Nagin D, Szczypula J. Comparative study of artificial neural network and statistical models for predicting student grade point averages. *Int J Forecast*. 1994;10(1):17-34.
9. Chipman ML, Steiner JW. An attempted simulation by computer of a committee on admissions. *Can Med Assoc J*. 1967;96(13):936-941.
10. Julian ER. Validity of the Medical College Admission Test for predicting medical school performance. *Acad Med*. 2005;80(10):910-917.
11. Reede JY. Predictors of success in medicine. *Clin Orthop Relat Res.* 1999;(362):72-77.
12. Silver B, Hodgson CS. Evaluating GPAs and MCAT scores as predictors of NBME I and clerkship performances based on students' data from one undergraduate institution. *Acad Med*. 1997;72(5):394-396.
13. Peskun C, Detsky A, Shandling M. Effectiveness of medical school admissions criteria in predicting residency ranking four years later. *Med Educ*. 2007;41(1):57-64.
14. Kidd RS, Latif DA. Traditional and novel predictors of classroom and clerkship success of pharmacy students. *Am J Pharm Educ*. 2003;67(4):Article 109.
15. Thomas MC, Draugalis JR. Utility of the Pharmacy College Admission Test (PCAT): implications for admissions committees. *Am J Pharm Educ*. 2001;66(1):47-51.
16. Smith FL, Grace SM, Kell C. Use of non-traditional assessments to reveal the relative value of PCAT versus GPA for admissions. *Am J Pharm Educ*. 2009;73(4):Article 57.
17. Roth MT, Mumper RJ, Singleton SF, et al. A renaissance in pharmacy education at the University of North Carolina at Chapel Hill. *N C Med J*. 2012;75(1):48-52.
18. US Department of Education Institute of Education Services National Center for Education Statistics. http://nces.ed.gov/ipeds/datacenter. Accessed May 25, 2015.
19. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.

20. Polishchuk PG, Muratov EN, Artemenko AG, Kolumbin OG, Muratov NN, Kuz'min VE. Application of random forest approach to QSAR prediction of aquatic toxicity. *J Chem Inf Model*. 2009;49(11): 2481-2488.

21. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont, CA: Wadsworth Publishing; 1984:358.

22. Kovdienko NA, Polishchuk PG, Muratov EN, et al. Application of random forest and multiple linear regression techniques to QSPR prediction of an aqueous solubility for military compounds. *Mol Inform*. 2010;29(5):394-406.

23. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model*. 2010;50(7): 1189-1204.

24. Fourches D, Muratov E, Tropsha A. Curation of chemogenomics data. *Nat Chem Biol.* 2015;11(8):535.

25. Meagher DG, Lin A, Stellato CP. A predictive validity study of the Pharmacy College Admission Test. *Am J Pharm Educ*. 2006;70(3): Article 53.

26. Meagher DG, Pan T, Perez CD. Predicting performance in the first-year of pharmacy school. *Am J Pharm Educ*. 2011;75(5): Article 81.

27. McCall KL, Allen DD, Fike DS. Predictors of academic success in a doctor of pharmacy program. *Am J Pharm Educ*. 2006;70(5): Article 106.