Predicting Step 2-CK performance using a machine learning approach

Padraig Healy, Syed Latifi

Submitted to: JMIR Medical Education
on: October 21, 2024

Table of Contents

Predicting Step 2-CK performance using a machine learning approach

Padraig Healy[1*] MSc; Syed Latifi[2*] PhD

[1]Weill Cornell Medicine- Qatar Doha QA
[2]Weill Cornell Medicine- Qatar Qatar Foundation - Education City Doha QA
[*]these authors contributed equally

Corresponding Author:
Padraig Healy MSc
Weill Cornell Medicine- Qatar
Qatar Foundation - Education City
Qatar Foundation
Doha
QA

Abstract

Background: In this study, we propose an innovative approach that leverages machine learning techniques to predict students' performance on United States Medical Licensing Examination (USMLE) Step-2 Clinical Knowledge (CK) exam. Our methodology involves the integration of feature selection and model validation processes within a nested cross-validation (CV) framework for Step-2 CK score prediction. Given the recent transition of the USMLE Step-1 exam to a pass/fail system, there is an anticipated shift in evaluative emphasis on the Step-2 Clinical Knowledge (CK) exam, prompting the need for advanced predictive models.

Methods: We conducted our analysis on data from four undergraduate medical student cohorts (Classes 2020 to 2023 inclusive, n = 117). A wide range of assessment data was considered by the algorithm for feature selection, including internal assessment data and National Board Medical Examination (NBME) Clinical Science Subject Exams scores. Utilizing nested cross-validation (CV), we constructed and assessed multiple regression models using four model evaluation metrics: mean CV error, adjusted-R2, Mallows Cp, and Bayesian Information Criteria (BIC).

Results: This led to the selection of a four-predictor model (adjusted-R2 = 0.68). This model incorporated a combination of NBME exams and performance in a pre-clinical unit.

Conclusion: Our approach effectively streamlines the process of building a predictive model by merging feature selection with model validation. By creating an interactive, user-friendly dashboard, we empower medical educators to predict students' Step-2 CK performance. This modeling and deployment approach holds promise for predicting student performances in other assessments.

(JMIR Preprints 21/10/2024:67776)
DOI: https://doi.org/10.2196/preprints.67776

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Predicting Step 2-CK performance using a machine learning approach

Healy P[1*], and Latifi S[1]

Affiliation(s):

1. *Division of Medical Education, Weill Cornell Medicine- Qatar, Qatar Foundation - Education City, P.O. Box 24144, Doha, Qatar*

***Corresponding Author**:

Padraig Mark Healy, MSc, Education Assessment Analyst, Office of Educational Development, Division of Medical Education, Weill Cornell Medicine- Qatar, Qatar Foundation - Education City, P.O. Box 24144, Doha, Qatar, pmh2003@qatar-med.cornell.edu

Authors' Information:
Padraig Mark Healy, MSc, Education Assessment Analyst, Office of Educational Development, Division of Medical Education. Orcid ID: 0000-0002-5804-0342
Syed Latifi, MSc, MEd, PhD, Acting Director, Office of Educational Development, Division of Medical Education. Orcid ID: 0000-0003-0505-609X

Abstract

Background: In this study, we propose an innovative approach that leverages machine learning techniques to predict students' performance on United States Medical Licensing Examination (USMLE) Step-2 Clinical Knowledge (CK) exam. Our methodology involves the integration of feature selection and model validation processes within a nested cross-validation (CV) framework for Step-2 CK score prediction. Given the recent transition of the USMLE Step-1 exam to a pass/fail system, there is an anticipated shift in evaluative emphasis on the Step-2 Clinical Knowledge (CK) exam, prompting the need for advanced predictive models.

Methods: We conducted our analysis on data from four undergraduate medical student cohorts (Classes 2020 to 2023 inclusive, n = 117). A wide range of assessment data was considered by the algorithm for feature selection, including internal assessment data and National Board Medical Examination (NBME) *Clinical Science Subject Exams* scores. Utilizing nested cross-validation (CV), we constructed and assessed multiple regression models using four model evaluation metrics: mean CV error, adjusted-R^2, Mallows Cp, and Bayesian Information Criteria (BIC).

Results: This led to the selection of a four-predictor model (adjusted-R^2 = 0.68). This model

incorporated a combination of NBME exams and performance in a pre-clinical unit.

Conclusion: Our approach effectively streamlines the process of building a predictive model by merging feature selection with model validation. By creating an interactive, user-friendly dashboard, we empower medical educators to predict students' Step-2 CK performance. This modeling and deployment approach holds promise for predicting student performances in other assessments.

Keywords: Nested cross-validation, machine learning, predictive model, USMLE Step-2 CK, interactive dashboard

Introduction

The United States Medical Licensing Examination (USMLE) is a three-step examination for medical licensure in the U.S., assessing competencies across, *Step-1*, *Step-2 Clinical Knowledge (CK)*, and *Step-3*.

Historically, Step-1 had a three-digit scoring system, and was often used by Residency Program Directors (RPDs) to shortlist candidates [1]. However, Step-1 transitioned to a pass/fail format in January 2022. Due to this change, the medical education community anticipates focus shifting towards Step-2 CK (which continues to provide a three-digit score) to objectively evaluate students' caliber and suitability for residency [2, 3, 4, 5, 6].

This expectation is supported by the National Resident Matching Program (NRMP)’s annual *Residency Program Director Survey*, which highlighted an increasing reliance on Step-2 CK scores as a critical component in the residency application assessment process. In 2022, this reliance peaked, with 96% of Residency Program Directors (RPDs) considering Step-2 CK scores informative when extending interview invitations to candidates from both U.S. medical graduates (USGs) and International Medical Graduates (IMGs) applicant groups [7,8]. In comparison to the previous years (2022 vs 2021), dependence on this rose by 2% among USGs and more pronounced 6% for IMGs. These data points not only underscore the integral role of the Step-2 CK in the residency selection landscape but also reflects the near-consensus among RPDs regarding its value as

a reliable and objective measure to shortlist the number of applications submitted to their program.

The significance of this trend was further amplified in the context of the 2024 NRMP Match, with 44,853 certified applicants (highest number on record) vying for one of 41,503 PGY-1 or PGY-2 training positions amongst the 6,395 certified programs on offer [9]. In such a competitive environment, the RPDs reliance on a quantifiable and objective metric like the Step-2 CK score has become naturally preferred, serving as a filter to manage the voluminous number of applications submitted to residency programs.

From the student's perspective, a recent survey of 4,649 prospective residency applicants, including US-MD, Doctor of Osteopathic Medicine (DO), and IMG students, found that knowing the minimum USMLE Step 2 or Comprehensive Osteopathic Medical Licensing Examination (COMLEX) Level 2 scores required by programs is highly desired [10]. This was especially important for IMG applicants, with 65.48% highlighting it as key, compared to 18.57% of US-MD and 15.90% of DO applicants (P<.001) [10]."

Given this context, there is a pressing need for advanced models to predict students' performance on the Step-2 CK exam. Such models can aid in early identification of students requiring academic support and aid advising efforts provided to NRMP applicants.

In this study, we propose an innovative approach leveraging machine learning techniques to predict students' performance on the USMLE Step-2 CK exam. Nested CV is a robust machine learning technique that combines model selection and validation, addressing the challenge of overfitting by using outer and inner folds for model selection and evaluation, respectively.

Methods

Traditionally, such a statistical modeling task would be conducted by partitioning a dataset into a training sample for model development and a test sample for evaluating predictive accuracy. A frequent deliberation within the research community revolves around determining 'an appropriate sample size,' which is crucial for maintaining the integrity and generalizability of the study's

findings. However, limited sample sizes often necessitate flexible and adaptive research methodologies. In response, we adhere to a commonly endorsed guideline in the scientific community, maintaining a minimum of 10-15 subjects per variable (SPV) to ensure the robustness of our model [11,12,13,14,15]. These considerations are particularly pertinent in medical schools with modest cohort size. Our program admits approximately 50 students per annum. Additionally, the extensive list of potential predictor variables, such as performances on preclinical units and National Board Medical Examination (NBME) exams, introduces the well-documented 'curse of dimensionality' [16] also known as '*big p, little n'* scenario. This occurs when a large number of predictor variables *(p)* contrasts with a small number of observations *(n)* posing a risk of overfitting. With insufficient observable data, the model may be unable to generalize and incorporate the noise, adapting too closely to the training dataset, but performing sub-optimally on unseen data.

Once we finalize a suitable model, we shift our focus to its deployment through an accessible, user-friendly dashboard tailored for stakeholders. This phase entails the implementation of a systematic approach consistent with the best practices in dashboard design, to ensure the resulting dashboard satisfies stakeholders requirements and achieves user-engagement.

This study was conducted at Weill Cornell Medicine – Qatar (WCM-Q). WCM-Q delivers a four-year medical degree program comprised of pre-clinical (years 1 and 2), and clinical years (years 3 and 4). Pre-clinical courses consist of units that are primarily assessed via quizzes. In the clinical years, core clerkships contain clinical, custom and NBME exam components. These assessments are listed in Table 1 below.

Table 1: Predictor variables considered during the feature selection stage.

Pre-clinical units (percent scores)	Biostatistics	Cardiovascular	Pulmonary	Gastrointestinal
	Renal	Hematology-Oncology	Brain and Behavior	Endocrinology
	Dermatology	Infectious	Rheumatology	Reproduction

	Diseases			
Clerkship NBME (Equated percent scores)	Medicine	Neurology	Surgery	Ob Gyn
	Pediatrics	Primary Care	Psychiatry	

Data Collection

Assessment performance records of Classes 2020 to 2023 inclusive were used in this study, for students who met the following inclusion criteria, i) sat for the new curriculum (implemented with Class 2020), ii) full performance vectors (pre-clinical units and NBME exams scores) were recorded, and iii) Step-2 CK scores were documented. For our analysis, we used the data from the students' first attempt at each assessment.

Data collection faced a constraint in the form of the introduction of a new curriculum, which restricted the sample size of this study, as the structure and sequence of unit delivery changed. This curriculum change provided a natural starting point for our study, ensuring (i) the collection of comparable data, and (ii) predictions made from the resulting model will be applicable to the curriculum in-situ going forward. This resulted in a total of 117 observations being used in the analysis.

Data Analysis

Stage-1: Feature Selection.

This stage involves selecting useful variables to explain the outcome while omitting the irrelevant ones. It is a process through which relevant features are retained for the machine learning model, aligning it with the intended modelling objectives. Statistical analyses were conducted using R language version 4.2.2 [17]. The function `regsubsets()` in `R [leaps package]` was utilized to select the best models based on the Residual Sum of Squares (RSS).

In this function, one specifies the maximum number of features to include in the model. We opted for a `nvar()`value of 5, which results in the algorithm returning five models: the best one-variable model, the best two-variable model, and so on, up to the best five-variable model. A maximum value

of 5 was taken to ensure that the largest model produced (5 variable), would maintain the 10-15 SPV rule of thumb [11,12,13,14,15]. The next step is to determine the overall optimal model using four model evaluation metrics.

Mean Cross-Fold Validation Error: It measures how well a model generalizes to unseen data, serving as an indicator of predictive accuracy. It is computed as the average squared error between predicted and observed values.

Adjusted R^2: It assesses how well a set of predictors explains the variation in the response variable while considering the number of predictors in model. A higher value indicates a better fit.

Mallows C_p: It compares the precision and bias of a model that includes all predictors to models with subsets of predictors. A lower value suggests a better model.

Bayesian Information Criteria (BIC): It measures test error and penalizes the addition of more variables to the model. A lower BIC value indicates a better model.

Stage-2: Cross Validation.

When working with a small sample-size, as is the case in this study, it is advisable to avoid the train-test approach, in which a portion of the data (usually 20-30%) is partitioned as test data [18,19]. Instead, an alternative approach known as cross-validation (CV) was used. The objective of cross-validation is equivalent to that of the train-test approach, which is to validate how well a predictive model, trained on seen data: a subset of the data used to train the model (training-set) will perform on unseen data: a subset of the data used to test the trained model (test-set). The difference lies in the cross-validation approach, where the data is divided into subsamples, and train-tests are conducted multiple times, efficiently utilizing the small data sample [18,19].

There are various types of cross validation, such as k-fold, stratified k-fold, and leave-one-out. In this application, we will use k-fold cross-validation. In this type, the value of k is a parameter chosen by the analyst. For instance, in our application using k=3 (indicating 3-fold cross-validation), we divide

the data into three equal parts. Due to the small sample size, three folds allocates a reasonable number of cases in each fold. We then run the validation process three times. In each iteration, two of the folds are used as the training-set, and one as the test-set. Next, models are fitted on the training data and evaluated on the test data. The accuracy of each iteration is computed, and the estimated error rate is derived as the average error rate from the three iterations.

Nested cross-validation. This approach combines Stage-1 and -2 and is recommended over the train-test approach in many instances [20, 21]. The order in which these two stages are executed is crucial. Applying the feature selection algorithm on the entire dataset (both train and test sets) before carrying out cross-validation can lead to positive bias, meaning the feature selection should occur nested within the cross-validation [22,23]. Having previously outlined each step individually, we next describe how these two processes combine to form a nested- or double cross-validation (see Figure 1).

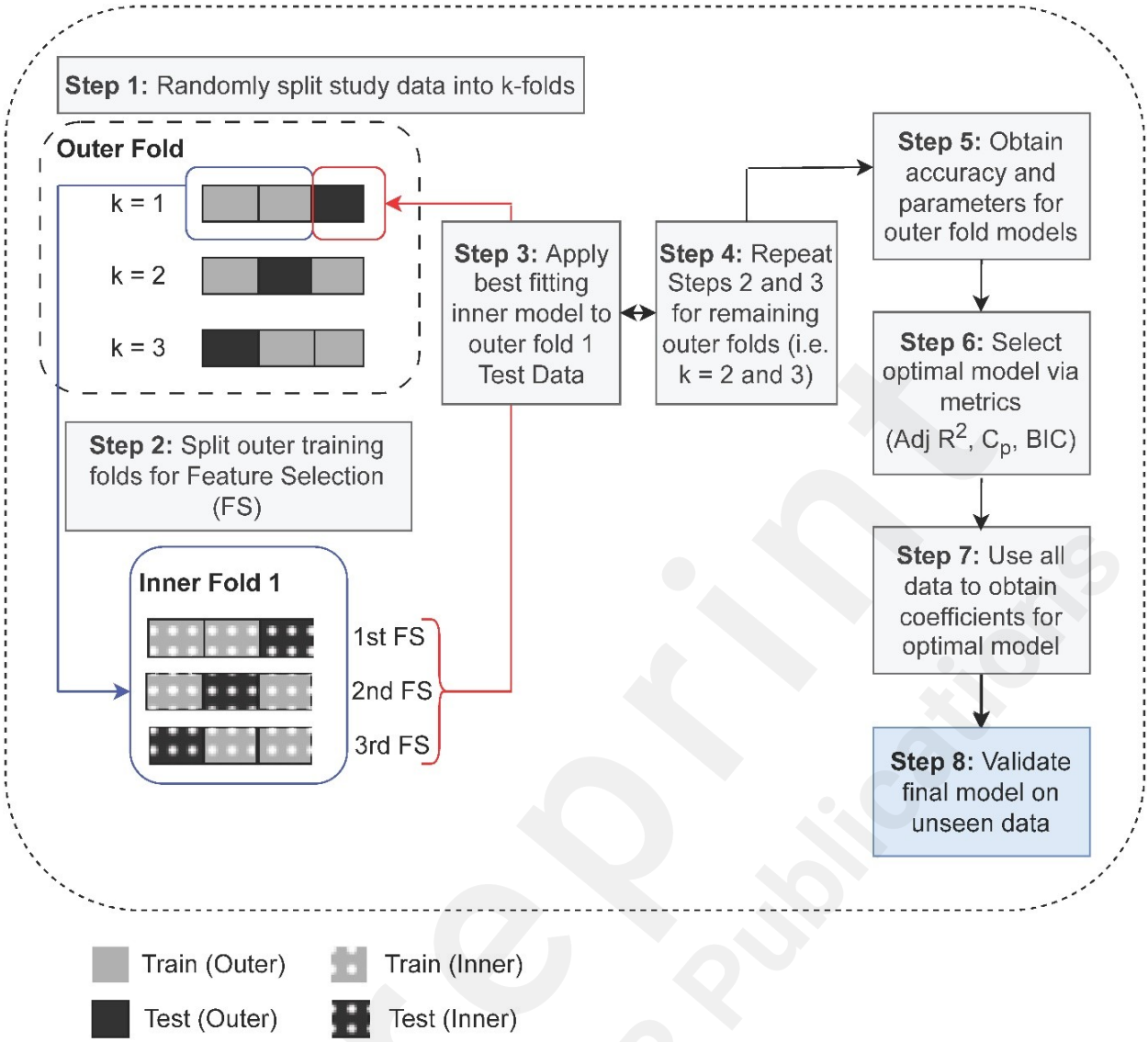

Figure 1: Nested cross-validation process

Step 1: Split the data into training and test data groupings (k = 3 in this application). **Step 2:** For each of the outer training folds, split them into inner folds for feature selection. **Step 3:** In this inner fold, select the best model and use it to test the outer fold. **Step 4:** Repeat steps 2 and 3 for the remaining outer folds. **Step 5:** Select the best outer models. **Step 6:** Select the optimal model based on metrics. **Step 7:** Use/apply the selected model to the entire dataset to obtain coefficients. **Step 8:** Validate the model on unseen data.

Results

A four-variable model emerges as the preferred choice for accurate performance prediction. To arrive at this optimal number of predictive variables we examined the following metrics: mean CV error, adjusted-R^2, Mallow's Cp, and BIC.

The mean CV error plot demonstrates how the predictive performance of a model changed with the

inclusion of different numbers of variables (Figure 2A). The most effective model, characterized by its peak predictive accuracy and minimal complexity, is indicated by the lowest point on this plot. Upon examining the mean CV error, we observed an improvement in the model's fit with the addition of the initial variables. However, after a certain number of variables were incorporated, the model's efficiency began to diminish. This reduction in efficiency became evident upon the inclusion of the fifth variable, i.e., the model reached an optimal level of efficiency at the fourth variable, after which further additions resulted in diminished performance.

Similarly, the adjusted-R^2 plot disclosed the balance between the models complexity and explanatory power – the variance explained by the predictors (Figure 2B). An higher adjusted-R^2 value signifies a more accurate model fit. We observed that the initial inclusion of more variables led to an increase in the adjusted-R^2 value, indicating a rise in the model's explanatory power. However, similar to the mean CV error plot, once the model exceeded a certain number of variables, the adjusted-R^2 value began to decline signifying that additional complexity no longer contributed to an increase in explanatory power.

Mallow's Cp and the BIC serve as model selection metrics that assess the trade-off between a model's fit to the data and its complexity (Figure 2C; Figure 2D). BIC, in particular, imposes a greater penalty for adding variables than Mallow's Cp, promoting a more judicious balance between fitting precision and simplicity. The ideal model, according to both metrics, is denoted by the lowest possible values. Our analysis revealed that initially adding variables reduced the values of both Mallow's Cp and BIC, indicating a model with a strong fit and low complexity. However, as more variables were included, these values began to rise, suggesting that any further complexity could lead to diminishing benefits and a risk of overfitting the model to the data. Overall, a four-variable model emerges as the preferred choice.

It optimizes the mean CV error and similarly minimizes values for Mallow's Cp and BIC. While the adjusted-R^2 peaks at 0.69 with a seven-predictor model, the four-predictor model achieves a

comparably high adjusted-R^2 of 0.68, making it the more parsimonious option which also aligns with the other three model evaluation criteria.

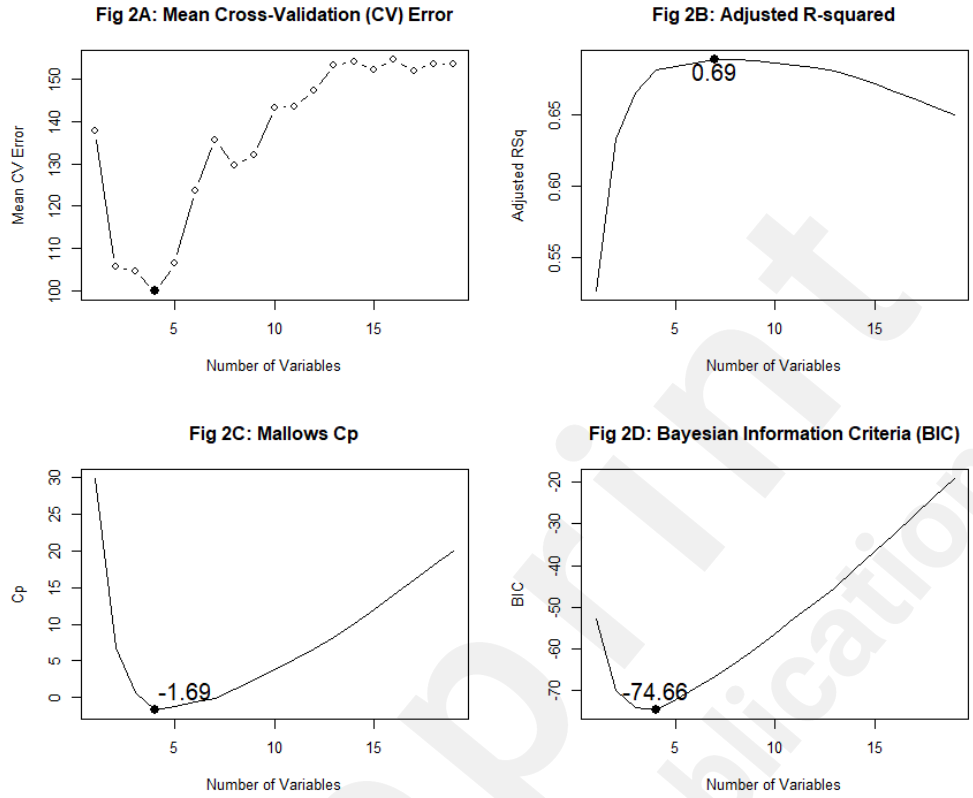

Figure 2: Metrics for selecting number of features in model: (A) Mean Cross-Validation (CV) Error, (B) Adjusted-R2, (C) Bayesian Information Criteria (BIC), (D) Mallows Cp.

Feature subset selection

Next, the algorithm investigates the optimal combination of variables (unit and NBME exam scores) needed to predict a student's performance on the Step-2 CK Exam. To determine the optimal features, the *best subset selection* technique was used. This consists of exhaustive selection algorithm, comparing all possible combination of predictor variables, and finds the best subset of variables according to the chosen criteria. The maximum feasible number of variables is the number of variables in the data frame. The resulting object produced by the `regsubsets()` command is a table displaying the best *n* feature variable models, as shown in Table 2.

Table 2: Subset selection of best n feature(s)

N predictor(	…	…	…	Gastrointestinal	…	…	…	…	…	Medicine	Neurology	Surgery

s)												
1	…	…	…	—	…	…	…	…	…	*	—	—
2	…	…	…	—	…	…	…	…	…	*	*	
3	…	…	…	—	…	…	…	…	…	*	*	*
4	…	…	…	*	…	…	…	…	…	*	*	*
…	…	…	…	…	…	…	…	…	…	…	…	…
19	*	*	*	*	*	*	*	*	*	*	*	*

* Indicates feature selected at the n^{th} predictor

With the model size decided upon, the next step is to obtain the coefficients. To do this, the best subset selection is conducted on the full data and the best four-variable model chosen. This resulted in three clerkship NBME exam first time performance being chosen by the model: Medicine, Neurology, Surgery, and one pre-clinical unit quiz performance: Gastrointestinal.

Model Comparison: Nested Cross-validation vs Train-Test

Next, we compared the nested CV model's performance in predicting outcomes to a more traditional modelling approach (train-test) using Mean Squared Error (MSE) values. For the train-test method, the data was randomly partitioned: 70% for training and 30% for testing. The same regression model was produced for both samples, predicted scores computed, and MSE's calculated.

The MSE values in Table 3 indicate the variability in prediction accuracy per different data samples and modelling methods. A lower MSE suggests that predicted values are close to observed value on average, indicating better model performance. The MSE obtained from nested CV (MSE = 99.90) falls between the MSEs from the Train and Test samples, demonstrating that the nested-CV approach has a performance level between that when used on train and test data. The lower MSE on training data (MSE = 91.79) is expected, as it indicates that the model performs well on the data it was trained on. In contrast, the MSE on the test set is the highest (MSE = 135.27), suggesting that the model struggled to generalize to unseen data by comparison.

Table 3: Comparison of model performance using MSE

	Nested-CV Approach	Train (70%)	Test (30%)
Mean Square Error	99.90	91.79	135.27

(MSE)

Dashboard Deployment

Having developed a predictive model, our focus now shifts toward its deployment. We contend that for a model to possess true utility, it must excel in two key areas: (i) predictive utility (i.e., how accurate it is at predicting scores) and (ii) end-user utility (i.e., ease of use and accessibility). End-user utility cannot be neglected; a complex model that is difficult for users to engage with, or too tedious to make a prediction with, will gain minimal adoption. To ensure our model's accessibility and ease of use, we leveraged Microsoft Power BI, a popular data visualization tool.

To initiate this process, we convened with stakeholders to establish and prioritize requirements. The objective of this dashboard was to facilitate the prediction of students' performances on the USMLE Step-2 CK exam. However, the dashboard designers sought additional information from stakeholders on what a fit-for-purpose dashboard should include from their perspectives as subject-matter experts. We suggested that they apply the MoSCoW prioritization framework [24] and designate all possible features into one of four categories: *'must have'*, *'should have'*, *'could have'* or *'will not have'*.

The design discussions for the dashboard's components led to the decision to incorporate five key features (see Figure 3a-e), each chosen for their specific utility in enhancing the predictive and user experience aspects of the tool (i.e., empowering users to interact with the model). Firstly, (a) a student name drop-down list for enabling retrieval of their scores, (b) input fields for the variables required to predict Step-2 CK performance, (c) predicted Step-2 CK score, (d) predicted quartile (program-level), and (e) a histogram displaying the distribution of historical students' first-time performances on the exam. The latter two features provide generalizable interpretability, i.e., enables the user to gauge the level of risk with a student by benchmarking them against the program's historical data (quartile and histogram). In addition to complementing the aforementioned measures, the histogram also provides a visual insight into the skewness and spread of student performances, as well as highlighting outliers. After eliciting this feedback, the next stage was *'storyboarding'*, in which the designers built a prototype to share with stakeholders for more feedback and refinement.

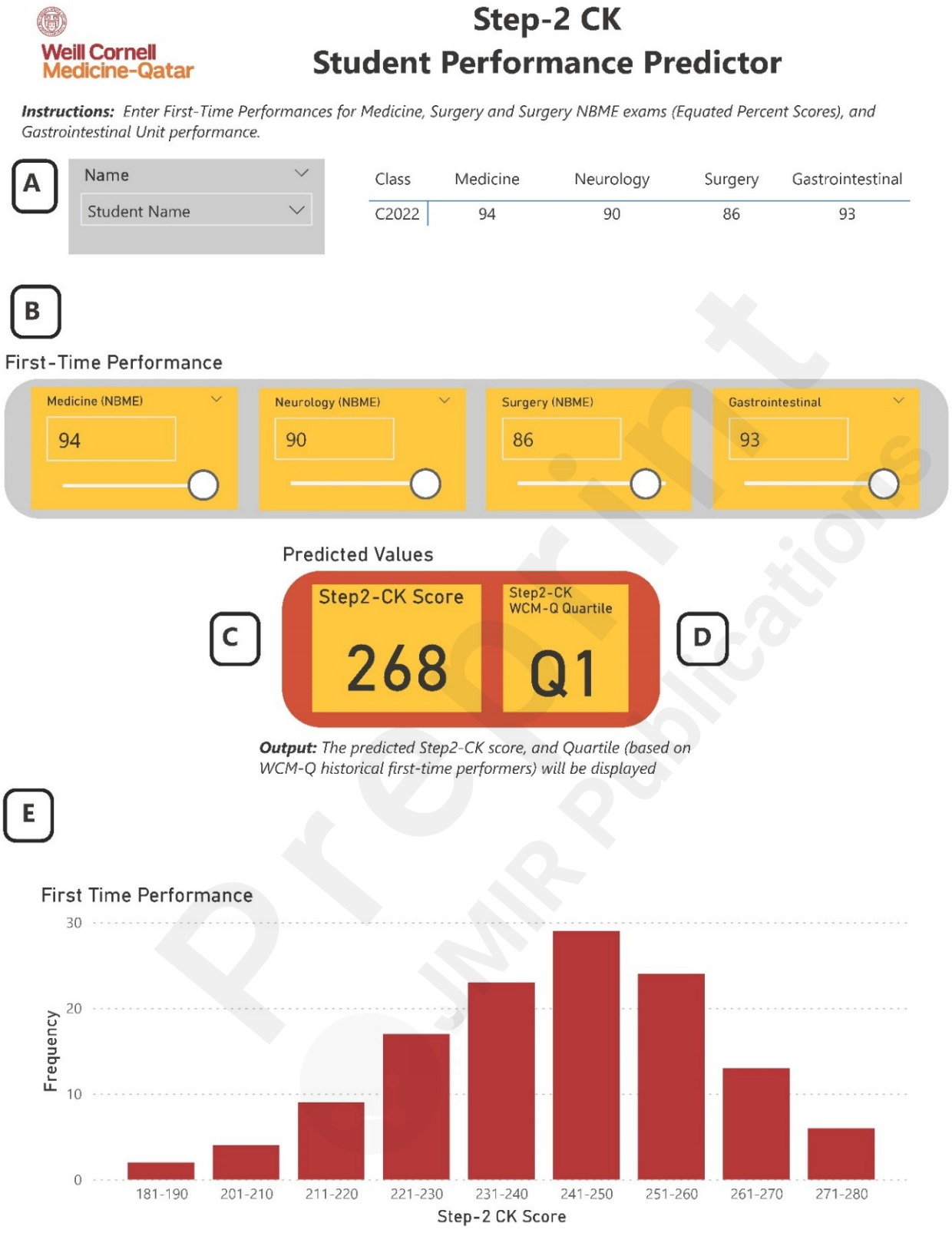

Figure 3: Step-2 CK Student Performance Predictor Dashboard

Discussion

Although the train-test split methodology has been a staple in machine learning for many years, it is not without its limitations [25]. This study introduces an alternative approach: nested cross-validation. This method offers advantages including a reduced propensity for overfitting and robustness even with small sample sizes, making it a preferred choice over the train-test split in many scenarios [20, 21]. We employed this approach to predict students' performance on the USMLE Step-2 CK Exam, utilizing historical preclinical units and NBME exams scores.

We have identified two primary use cases for this model. The first pertains to Residency Program Directors, who are increasingly factoring in Step-2 CK scores when extending interview invitations for their residency programs [2,3,4,5,6]. A model capable of predicting Step-2 CK scores could significantly aid medical educators in advising students on residency applications. Secondly, the model can also serve to aid identifying students who may be academically at risk, thereby facilitating timely intervention.

While the model has demonstrated robustness even with small sample sizes, one potential limitation of this study is the sample size. The specific characteristics of these cohorts may limit the generalizability of the results. Another limitation of the study is the lack of control for socio-demographic factors such as age, race, ethnicity, and social class, which could affect the generalizability of the results [26].

Future work will involve implementing a cyclical recalibration process for the algorithm when a new cohorts' data is incorporated into the database. This measure is designed to mitigate a phenomenon known as model drift, where the model's predictive power diminishes over time due to shifts in the underlying data. Furthermore, this methodology could be extended to construct and deploy similar predictive models for other high-stakes assessments within the educational program.

Conclusion

This study illustrates (i) how machine learning approaches can be applied in medical education, and (ii) how such a model can be deployed for use by the medical educator via an interactive dashboard. This work explains the development of a machine-learning model, aiming to maximize predictive- and user- utility for tasks such as early identification of students who may face challenges with the Step-2 CK exam. This allows for a timely intervention to assist and guide students with suitable residency applications.

List of Abbreviations

BIC: Bayesian Information Criteria
COMLEX: Comprehensive Osteopathic Medical Licensing Examination
CV: Cross-validation
DO: Doctor of Osteopathic Medicine
IMGs: International Medical Graduates
MSE: Mean Squared Error
NRMP: National Resident Matching Program
PG: Postgraduate
RPD: Residency Program Director
RSS: Residual Sum of Squares
SPV: subjects per variable
Step-2 CK: Step-2 Clinical Knowledge
USGs: United States Medical Graduates
USMLE: United States Medical Licensing Examination

Acknowledgements

The authors would like to recognize and appreciate Thurayya Arayssi M.D., FACP, FACR, FRCP, Vice Dean for Academic and Curricular Affairs, Professor of Clinical Medicine, Division of Medical Education, Weill Cornell Medicine-Qatar, for her insightful comments and suggestions, which enhanced the quality of this manuscript

The authors would also like to recognize and appreciate Dr. Philippe Piccardi, Specialist, Scientific and Education Content, Weill Cornell Medicine-Qatar, who helped with reviewing this manuscript.

Funding

The authors did not receive any research funding for this project.

Availability of data and material

The data used in this study are available from the corresponding author upon a reasonable request.

Declaration of interest statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Declarations

Ethics approval and consent to participate

The study was reviewed by the Office of the Institutional Review Board (IRB) at Weill Cornell Medicine - Qatar, which determined that this study is a quality improvement, non-research activity.

Therefore, ethics approval and consent were not required.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Green M, Jones P, Thomas Jr JX. Selection criteria for residency: results of a national program directors survey. Academic Medicine. 2009 Mar 1;84(3):362-7.
2. Bird JB, Olvet DM, Willey JM, Brenner JM. A Generalizable Approach to Predicting Performance on USMLE Step 2 CK. Advances in Medical Education and Practice. 2022;13:939.
3. Chisholm LP, Drolet BC. USMLE Step 1 scoring changes and the urology residency application process: program directors' perspectives. Urology. 2020 Nov 1;145:79-82.
4. MacKinnon GE, Payne S, Drolet BC, Motuzas C. Pass/Fail USMLE Step 1 Scoring—a radiology program director survey. Academic Radiology. 2021 Nov 1;28(11):1622-5.
5. Pontell ME, Makhoul AT, Kumar NG, Drolet BC. The change of USMLE step 1 to pass/fail: perspectives of the surgery program director. Journal of surgical education. 2021 Jan 1;78(1):91-8.
6. Ozair A, Bhat V, Detchou DK. The US residency selection process after the United States Medical Licensing Examination Step 1 pass/fail change: overview for applicants and educators. JMIR Medical Education. 2023 Jan 6;9(1):e37069.
7. National Resident Matching Program. Data Release and Research Committee. *Results of the 2021 NRMP Program Director Survey*. National Resident Matching Program, Data Release and Research Committee. 2021. https://www.nrmp.org/wp-content/uploads/2021/11/2021-PD-Survey-Report-for-WWW.pdf. Accessed July 18, 2024.
8. National Resident Matching Program. Data Release and Research Committee. *Results of the 2022 NRMP Program Director Survey*. National Resident Matching Program, Data Release and Research Committee. 2022. https://www.nrmp.org/wp-content/uploads/2022/09/PD-Survey-Report-2022_FINALrev.pdf. Accessed July 18, 2024.
9. National Resident Matching Program. Advance data tables: 2024 main residency match. 2024. https://www.nrmp.org/wp-content/uploads/2024/03/Advance-Data-Tables-2024.pdf. Accessed July 18, 2024.
10. Ulin L, Bernstein SA, Nunes JC, Gu A, Hammoud MM, Gold JA, Mirza KM. Improving transparency in the residency application process: survey study. JMIR Formative Research. 2023 Dec 25;7:e45919.
11. Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. Journal of clinical epidemiology. 2015 Jun 1;68(6):627-36.
12. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic medicine. 2004 May 1;66(3):411-21.
13. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: springer; 2001 Jan 10.
14. Schmidt FL. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. Educational and Psychological Measurement. 1971 Oct;31(3):699-714.
15. VanVoorhis CW, Morgan BL. Understanding power and rules of thumb for determining sample sizes. Tutorials in quantitative methods for psychology. 2007 Sep 1;3(2):43-50.
16. Altman N, Krzywinski M. The curse (s) of dimensionality. Nature Methods. 2018 Jun 1;15(6):399-400.
17. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2022. https://www.R-project.org. Accessed July 18, 2024.
18. Beleites C, Baumgartner R, Bowman C, Somorjai R, Steiner G, Salzer R, Sowa MG. Variance reduction in estimating classification error using sparse datasets. Chemometrics and intelligent laboratory systems. 2005 Oct 28;79(1-2):91-100.
19. Hawkins DM, Basak SC, Mills D. Assessing model fit by cross-validation. Journal of chemical

information and computer sciences. 2003 Mar 24;43(2):579-86.
20. Lewis MJ, Spiliopoulou A, Goldmann K, Pitzalis C, McKeigue P, Barnes MR. nestedcv: an R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data. Bioinformatics Advances. 2023 Jan 1;3(1):vbad048.
21. De Rooij M, Weeda W. Cross-validation: A method every psychologist should know. Advances in Methods and Practices in Psychological Science. 2020 Jun;3(2):248-63.
22. Demircioğlu A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights into Imaging. 2021 Dec;12:1-0.
23. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: springer; 2013 Jun 24.
24. Hatton S. Early prioritisation of goals. In Advances in Conceptual Modeling–Foundations and Applications: ER 2007 Workshops CMLSA, FP-UML, ONISW, QoIS, RIGiM, SeCoGIS, Auckland, New Zealand, November 5-9, 2007. Proceedings 26 2007 (pp. 235-244). Springer Berlin Heidelberg.
25. Singh V, Pencina M, Einstein AJ, Liang JX, Berman DS, Slomka P. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. Scientific reports. 2021 Jul 14;11(1):14490.
26. Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, 75(3), 417-453.