BIOS6621: Final Project

COMPREHENSIVE REPORT

Project title: OSCE scores, URM status, and USMLE performance among medical students

Submitted to: Dr. Todd Guth, School of Medicine, University of Colorado Anschutz Medical Campus.

Report prepared by: Mahfuza Haque Mahi

Date: 30-11-2025

Introduction

Objective Structured Clinical Examination (OSCE) scores are an essential tool for evaluating medical students' clinical skills, including communication, documentation, and physical examination. While many prior studies have evaluated individual OSCE components, relatively few have simultaneously examined multiple OSCE domains or compared performance across student subgroups defined by underrepresented in medicine (URM) status or gender.

In addition, standardized examinations such as USMLE Step 1 and Step 2 CK remain key milestones in medical education and may reflect or differ from clinical performance measured by OSCEs. Understanding how OSCE performance relates to USMLE scores, and whether these relationships differ by URM status, can inform efforts to promote equity in assessment and support students at risk of differential evaluation.

This analysis focuses on a single medical school cohort of third-year students and addresses the following research questions:

Primary hypothesis:

- H1: Mean OSCE domain scores (communication, documentation, physical exam) differ between URM and non-URM students.

Secondary hypotheses:

- H2: Mean USMLE Step 1 and Step 2 CK scores differ between URM and non-URM students.
- H3: OSCE domain scores are positively correlated with USMLE Step examination scores.

A cross-sectional analysis was conducted using student-level summaries of OSCE scores across multiple pre-clinical OSCE sessions, along with USMLE Step scores and URM status.

Methods

Study design

This is a cross-sectional analysis of a single cohort of third-year medical students from one medical school.

Each student completed multiple OSCE sessions during the pre-clinical curriculum which produced several domain-specific scores per student. Because our goal was to describe overall performance rather than time trends, we averaged each domain (Communication, Documentation and Physical examination) across all available sessions. This created a single, interpretable summary score per domain and avoided unnecessary modeling complexity.

USMLE Step 1 and Step 2 CK scores were linked at the student level.

Study population

The original dataset included 199 third-year medical students. For each student, multiple OSCE scores were recorded across pre-clinical courses (e.g., Bates, Ramos, Payne, Hyde, Snelling). A set of clerkship clinical and final grades (e.g., AAC, RCC) was also available but had substantial missingness.

For the present analysis, we restricted to students with at least one non-missing OSCE domain score, resulting in an analytic sample of 160 students (approximately 80.4% of the original cohort).

Data

Data Sources and Structure

Data were obtained from institutional academic records for the third-year medical students. These records included information across several domains:

Demographics: URM indicator (URM), Alternative URM coding (URM_AAMC), Gender (gender), Race (race_desc).

OSCE performance: OSCE scores from multiple pre-clinical courses and sessions, with domain-specific scores for communication, physical exam, and medical documentation (e.g., P1SprBates_CommScore, P1SprBates_PhysExamScore, P1SprBates_MedDocScore, P2FallRamos_CommScore, P2FallRamos_PhysExamScore, etc.).

USMLE scores: Student-level scores for the USMLE Step 1 and Step 2 CK examinations (STEP1_ExamScore, STEP2CK_ExamScore).

Clerkship grades (clinical and final): AAC_clinical_grade, AAC_Final_Course_Grade, RCC_clinical_grade, RCC_Final_Course_Grade, etc.

Data were collected during the pre-clinical phases of medical school. OSCE scores were obtained from course-level assessments, and USMLE scores from NBME-reported data. Each student completed OSCEs across several courses in the pre-clinical phase (e.g., Bates, Ramos, Payne, Hyde, Snelling). Because OSCEs were assessed multiple times per student across sessions, the raw dataset contained several score columns for each of the three OSCE domains.

Data quality and pre-processing

Several data quality checks were conducted prior to analysis. Clinical and final clerkship grades were validated against the allowable coding scheme (0, 1, 2, 3, and 9), and no invalid entries were identified. Gender was predominantly coded as “F” or “M,” with a single irregular value (“D”), which was recoded as missing.

Race categories were merged to address duplicated, overlapping, and extremely small subgroups in the raw data. Similar labels (e.g., “African,” “African American,” “Black or African American”) were combined into broader, conceptually consistent categories. Very small or infrequent categories were collapsed into an “Other” group to ensure stable, interpretable summaries.

Consistency between URM and URM_AAMC classifications was examined using cross-tabulations. A small number of discrepancies were identified (e.g., cases coded as URM = 1 but URM_AAMC = 0). For simplicity and to maintain consistency with the prespecified analytic plan, a binary URM variable was derived directly from URM (1 = URM, 0 = non-URM).

Missingness patterns were also evaluated. Several clerkship grade variables (e.g., AAC and RCC clinical/final grades) exhibited more than 85% missingness, and one ethnicity indicator (amcas_othethnicity) was entirely missing. These variables were excluded from further analysis. Additionally, 39 of the 199 students (approximately 19.6%)) were missing all OSCE domain scores and were excluded from analyses involving OSCE outcomes.

Statistical methods

Descriptive analysis

Table 1 summarizes demographics (URM status, gender and race) and key continuous variables (OSCE domain means and USMLE Step scores), stratified by URM status. The overall distributions of OSCE domain scores and USMLE Step scores were evaluated using histograms, whereas stratified boxplots were used to compare OSCE and USMLE scores between URM and non-URM students.

Primary analysis (H1: OSCE domain differences by URM)

1. Unadjusted comparisons: Two-sample Welch t-tests were conducted to compare mean OSCE Communication, Documentation, and Physical Exam scores between URM and non-URM students. This method is appropriate when comparing the mean of a continuous score between two groups. It tells us whether any observed differences are larger than what we would expect by chance alone.
2. Adjusted models: For each OSCE domain, a linear regression model was fit to account for potential confounding factors :

$$\text{OSCE Domain Mean}_i = \beta_0 + \beta_1 URM_i + \beta_2 Gender_i + \beta_3 Step1_i + \beta_4 Step2_i + \epsilon_i$$

where β_1 represents the adjusted mean difference in OSCE scores between URM and non-URM students.

These regregression models estimated the adjusted association between URM status and OSCE performance while controlling for demographic and academic factors.

3. Multiple comparison adjustment: To account for testing across the three OSCE domains, false discovery rate (FDR) correction was applied to the three URM-domain p-values. As we evaluated three OSCE domains and testing multiple outcomes increases the chance of finding a “significant” result by accident. The False Discovery Rate (FDR) adjustment helps control this risk and makes our conclusions more reliable.

Secondary Analysis (H2): Differences in USMLE Scores by URM Status

1. Unadjusted comparisons: USMLE Step 1 and Step 2 CK scores were compared between URM and non-URM students using Welch’s t-tests.
2. Adjusted models: For each USMLE exam, a linear regression model was fit adjusting for gender, providing adjusted estimates of score differences between URM and non-URM students.

Association Analysis (H3): Relationship Between OSCE and USMLE Scores

1. Correlation analysis: Pearson correlation coefficients were computed between each of the three OSCE domain means and the two USMLE exam scores. These correlations quantified the direction and strength of linear associations between OSCE scores and USMLE scores.
2. Simple linear regression: For each OSCE–USMLE pair, a simple linear regression model was fit with:

Outcome: USMLE Step score

Predictor: OSCE domain mean

These models assessed how much change in USMLE performance was associated with a one-unit increase in OSCE performance.

Software

All analyses were conducted in R (R version 4.3.1).

Results

Descriptive statistics (Table 1)

Table 1 below summarizes demographic characteristics and assessment outcomes (OSCE and USMLE scores) for the cohort, stratified by URM status.

Table 1: Baseline demographic characteristics and assessment outcomes (OSCE domain scores and USMLE Step scores) among medical students, stratified by URM status.

	Non-URM	URM
	(N=112)	(N=48)
Gender		
F	58 (51.8%)	22 (45.8%)
M	53 (47.3%)	26 (54.2%)
Missing	1 (0.9%)	0 (0%)
Race		
White	85 (75.9%)	11 (22.9%)
Asian	15 (13.4%)	10 (20.8%)
Black / African American	0 (0%)	10 (20.8%)
American Indian / Alaska Native	0 (0%)	7 (14.6%)
Other	4 (3.6%)	2 (4.2%)
Missing	8 (7.1%)	8 (16.7%)
USMLE Step 1		
Mean (SD)	230 (21.0)	220 (17.8)
Median [Min, Max]	232 [164, 267]	221 [189, 259]
USMLE Step 2 CK		
Mean (SD)	247 (13.1)	238 (12.9)
Median [Min, Max]	248 [209, 276]	239 [211, 262]
Missing	0 (0%)	1 (2.1%)
OSCE Communication (%)		
Mean (SD)	92.6 (4.65)	91.6 (4.60)
Median [Min, Max]	93.6 [76.3, 100]	92.4 [81.4, 98.3]
OSCE Documentation (%)		
Mean (SD)	92.3 (4.67)	90.4 (5.77)
Median [Min, Max]	93.1 [74.6, 100]	90.6 [72.9, 100]
OSCE Physical Exam (%)		
Mean (SD)	90.5 (4.00)	88.7 (5.33)
Median [Min, Max]	90.8 [77.4, 98.0]	89.5 [74.3, 96.7]

Among the 160 students in the analytic sample, 112 (70%) were classified as Non-URM and 48 (30%) as URM.

Gender distribution was similar across groups. Among Non-URM students, 51.8% were female and 47.3% were male. Among URM students, 45.8% were female and 54.2% were male.

Race distributions differed substantially between URM and non-URM students. Among non-URM students, the majority identified as White (75.9%), followed by Asian (13.4%). Among URM students, White (22.9%) and Asian (20.8%) students were less common, with higher representation in Black/African American (20.8%) and American Indian/Alaska Native (14.6%) categories. Missing race information was more common among URM students (16.7%) compared with non-URM students (7.1%).

Mean USMLE scores were higher among Non-URM students for both Step examinations. Step 1 scores averaged 229.8 (SD 21) for Non-URM students and 219.9 (SD 17.8) for URM students. A similar pattern was observed for Step 2 CK scores (Non-URM mean 246.6, SD 13.1; URM mean 238.2, SD 12.9).

OSCE domain scores were high overall, showing expected ceiling effects. Mean OSCE Communication scores were 92.6% for Non-URM and 91.6% for URM students. Documentation scores averaged 92.3% (Non-URM) and 90.4% (URM). Physical Exam scores were slightly lower overall but still high, averaging 90.5% among Non-URM and 88.7% among URM students.

Exploratory data visualization

Histograms of USMLE Step 1 and Step 2 CK scores (Appendix Figure 1) show approximately normal distributions with moderate variability and no extreme outliers. Histograms of OSCE Communication, Documentation, and Physical Exam scores (Appendix Figure 2) display pronounced right-skew and ceiling effects, with most students scoring between 85–100%. Boxplots of USMLE Step scores by URM status (Appendix Figure 3) illustrate that URM students tended to score lower on both Step 1 and Step 2 CK than non-URM students, with visibly lower medians and similar variability across groups. Boxplots of OSCE domain scores by URM status (Appendix Figure 4) show high scores and substantial overlap between groups, with only small visual differences between URM and non-URM students and strong ceiling effects across all domains.

Primary analysis: OSCE domain differences by URM (H1)

Table 2 below summarizes the unadjusted mean differences in OSCE domain scores between URM and Non-URM students.

Table 2: Two-sample t-tests comparing mean OSCE domain scores between URM and Non-URM students.

Domain	Mean Non-URM	Mean URM	Mean Difference	95% CI	p-value
Communication	92.64	91.64	1.00	-0.58 to 2.58	0.213
Documentation	92.35	90.44	1.91	0.03 to 3.79	0.046
Physical Exam	90.54	88.65	1.89	0.18 to 3.6	0.031

In unadjusted analyses, Non-URM students scored slighlty higher than URM students across all OSCE domains. The difference in Communication scores was small and not statistically significant (mean difference = 1, 95% CI: -0.58 to 2.58; p = 0.21). In contrast, Non-URM students scored significantly higher in both Documentation (mean difference = 1.91, 95% CI: 0.03 to 3.79; p = 0.046) and Physical Exam performance (mean difference = 1.89, 95% CI:0.18 to 3.6; p = 0.031). These results suggest modest but statistically significant differences in two of the three OSCE domains prior to adjustment for covariates.

Table 3 presents the results of adjusted linear regression models examining the associations of URM status, gender, and USMLE scores with OSCE domain performance.

Table 3: Adjusted linear regression models for OSCE domain scores with URM status, gender, and USMLE scores as predictors.

Domain	Predictor	Estimate	95% CI	p-value
Communication	Intercept	73.48	59.99 to 86.97	<0.001
	Male (vs Female)	-2.13	-3.55 to -0.7	0.004
	URM (vs Non-URM)	-0.10	-1.7 to 1.49	0.899
	USMLE Step 1	0.00	-0.05 to 0.05	0.971
	USMLE Step 2 CK	0.08	0.01 to 0.16	0.033
Documentation	Intercept	70.81	56.11 to 85.52	<0.001
	Male (vs Female)	-0.86	-2.41 to 0.69	0.275
	URM (vs Non-URM)	-1.42	-3.16 to 0.32	0.109
	USMLE Step 1	-0.04	-0.1 to 0.01	0.130
	USMLE Step 2 CK	0.13	0.05 to 0.21	0.003
Physical Exam	Intercept	66.57	53.81 to 79.33	<0.001
	Male (vs Female)	-1.71	-3.06 to -0.37	0.013
	URM (vs Non-URM)	-0.94	-2.45 to 0.57	0.223
	USMLE Step 1	0.01	-0.04 to 0.05	0.825
	USMLE Step 2 CK	0.10	0.02 to 0.17	0.009

In the OSCE communication domain, the URM coefficient was essentially zero (estimate -0.1; p = 0.899), indicating no meaningful adjusted difference between groups. For Documentation, URM students scored an estimated -1.42 points lower on average, but this difference was not statistically significant (p = 0.109). Similarly, in the Physical Exam domain, the adjusted URM difference was less than 1 point and non-significant (p = 0.223). Across all domains, gender and USMLE Step 2 CK scores showed consistent associations with OSCE performance, whereas USMLE Step 1 did not meaningfully contribute. Overall, these adjusted models provide no evidence that URM status independently predicts OSCE domain scores.

The following Table 4 summarizes the FDR-corrected significance levels for URM effects across OSCE domains.

Table 4: False Discovery Rate (FDR)–adjusted p-values for URM effects across OSCE domains.

Domain	Raw_p_value	FDR_adjusted_p_value
Communication	0.899	0.899
Documentation	0.109	0.326
Physical Exam	0.223	0.334

After applying False Discovery Rate (FDR) correction to account for multiple comparisons across the three OSCE domains, none of the URM effects remained statistically significant. Although unadjusted analyses suggested slightly lower Documentation and Physical Exam scores among URM students, these differences did not persist after correction. Overall, the FDR-adjusted results provide no evidence of differences in OSCE domain performance by URM status.

Secondary analysis: USMLE differences by URM (H2)

The following table presents the unadjusted comparisons of USMLE Step scores between Non-URM and URM students.

Table 5: Two-sample t-tests comparing USMLE Step scores between Non-URM and URM students.

Outcome	Mean Non-URM	Mean URM	Mean Difference (Non-URM – URM)	95% CI	p-value
USMLE Step 1	229.8	219.9	9.85	3.42 to 16.28	0.003
USMLE Step 2 CK	246.6	238.2	8.37	3.9 to 12.84	<0.001

In unadjusted comparisons, URM students scored significantly lower than Non-URM students on both USMLE Step examinations. For Step 1, Non-URM students scored approximately 10 points higher on average (229.8 vs. 219.9; p = 0.003). A similar pattern was observed for Step 2 CK, where Non-URM students scored about 8 points higher on average (246.6 vs. 238.2; p < 0.001). These unadjusted results indicate meaningful differences in USMLE performance between URM and non-URM students prior to adjustment for other factors.

Table 6 below displays gender-adjusted differences in USMLE Step scores between URM and Non-URM groups.

Table 6: Adjusted differences in USMLE Step scores between URM and Non-URM students.

Outcome	Adjusted Difference (URM – Non-URM)	95% CI	p-value
USMLE Step 1	-10.1	-17 to -3.2	0.004
USMLE Step 2 CK	-8.3	-12.8 to -3.8	<0.001

After adjusting for gender, URM students continued to score significantly lower on both Step examinations. URM students scored an estimated -10.1 points lower on Step 1 (95% CI: -17 to -3.2; p = 0.004). For Step 2 CK, URM students scored -8.3 points lower (95% CI: -12.8 to -3.8; p = <0.001). These adjusted differences remained statistically significant and suggest that URM students experience notable performance gaps on standardized licensing examinations that are not explained by gender alone.

OSCE-USMLE relationship (H3)

The following Table 7 summarizes the strength and significance of associations between OSCE domain scores and USMLE Step examination scores.

Table 7: Pearson correlations between OSCE domain scores and USMLE Step scores.

OSCE Domain	USMLE Exam	r	95% CI	p-value
OSCE Communication	USMLE Step 1	0.15	[-0.01, 0.30]	0.062
OSCE Documentation	USMLE Step 1	0.11	[-0.05, 0.26]	0.165
OSCE Physical Exam	USMLE Step 1	0.24	[0.08, 0.38]	0.003
OSCE Communication	USMLE Step 2 CK	0.26	[0.11, 0.40]	0.001
OSCE Documentation	USMLE Step 2 CK	0.26	[0.11, 0.40]	<0.001
OSCE Physical Exam	USMLE Step 2 CK	0.34	[0.20, 0.48]	<0.001

All OSCE domains showed positive correlations with USMLE Step performance. Associations were consistently stronger for Step 2 CK than for Step 1, with Step 2 CK correlations falling in the modest but statistically significant range. For USMLE Step 1 score, OSCE Communication and Documentation showed weak, non-significant correlations (r = 0.15 and r = 0.11, respectively), whereas Physical Exam showed a small but statistically significant association (r = 0.24; p = 0.003). For USMLE Step 2 CK score, all three OSCE domains showed statistically significant correlations, ranging from r = 0.26 to r = 0.34, indicating modest but meaningful associations. Overall, these findings suggest that OSCE performance is more closely aligned with the applied clinical knowledge assessed by Step 2 CK than with the foundational science content assessed in Step 1.

This table below shows the associations between OSCE domain scores and USMLE Step scores estimated from simple linear regression models.

Table 8: Associations between OSCE domain scores and USMLE Step scores from simple linear regression models.

Outcome	Predictor	Estimate	CI	p-value
USMLE Step 1	OSCE Communication	0.655	[-0.033, 1.342]	0.062
	OSCE Documentation	0.446	[-0.185, 1.078]	0.165
	OSCE Physical Exam	1.080	[0.384, 1.777]	0.003
USMLE Step 2 CK	OSCE Communication	0.750	[0.307, 1.193]	0.001
	OSCE Documentation	0.704	[0.298, 1.109]	0.001
	OSCE Physical Exam	1.033	[0.59, 1.476]	0.000

Across all models, higher OSCE domain scores were associated with higher USMLE Step scores. For USMLE Step 1, only the OSCE Physical Exam domain showed a statistically significant positive association (β = 1.08, 95% CI [0.384, 1.777]; p = 0.003). OSCE Communication and Documentation domain scores showed positive but non-significant associations with Step 1 (β = 0.655 and β = 0.446, respectively).

For USMLE Step 2 CK, all three OSCE domains were significantly associated with exam performance. The strongest relationship was observed for Physical Exam (β = 1.033, 95% CI [0.59, 1.476]; p = 0.000). Communication (β = 0.75) and Documentation (β = 0.704) also showed statistically significant associations.

Overall, OSCE performance, particularly in the Physical Exam domain shows meaningful associations with USMLE scores, with stronger relationships observed for Step 2 CK than Step 1.

Discussion

In this cohort of medical students, OSCE domain scores were uniformly high, demonstrating clear ceiling effects in communication, documentation, and physical exam performance. Although non-URM students showed slightly higher unadjusted scores in the documentation and physical exam domains, these differences were not statistically significant after adjustment for gender and USMLE scores or after correction for multiple comparisons.

In contrast, substantial group differences were observed in standardized examination performance. Non-URM students scored significantly higher than URM students on both USMLE Step 1 and Step 2 CK examinations, and these differences remained after adjustments for gender. These findings suggest that performance gaps were more evident on written licensing examinations than on OSCE-based clinical skills assessments in this cohort.

Modest positive associations were observed between OSCE domain scores and USMLE scores, particularly for Step 2 CK. However, the strength of these correlations indicates that OSCEs and USMLE exams assess related but distinct components of clinical competence, with OSCEs emphasizing communication, documentation, and physical examination skills and USMLE exams assessing foundational knowledge and clinical reasoning. Together, these findings highlight the multifaceted nature of clinical competency assessment and the importance of using multiple measures to evaluate student performance.

Limitations

This study has several limitations. The analysis was conducted within a single institution and class cohort, which may limit the generalizability of the findings to other medical schools. OSCE domain scores exhibited clear ceiling effects, reducing the ability to detect subtle performance differences between groups. Approximately 20% of students were missing all OSCE domain scores and were therefore excluded. If these data were not missing completely at random, then selection bias may have influenced the results. The simplified binary URM indicator may not fully capture the heterogeneity of students' racial and ethnic identities or align perfectly with AAMC URM definitions. Finally, although adjusted models accounted for gender and USMLE scores, unmeasured confounding such as prior academic performance or socioeconomic background may have contributed to the observed associations.

Future directions

These findings suggest several directions for expanding the scope of future work. Extending the analysis to additional cohorts would allow investigation of temporal trends and improve the precision of estimates. Evaluating potential gender-by-URM interactions and other intersectional patterns may provide a more nuanced understanding of performance patterns across student subgroups. Furthermore, incorporating item-level or station-level OSCE data could help identify whether particular skills or content areas exhibit larger performance gaps than overall domain averages.

Figures Appendix

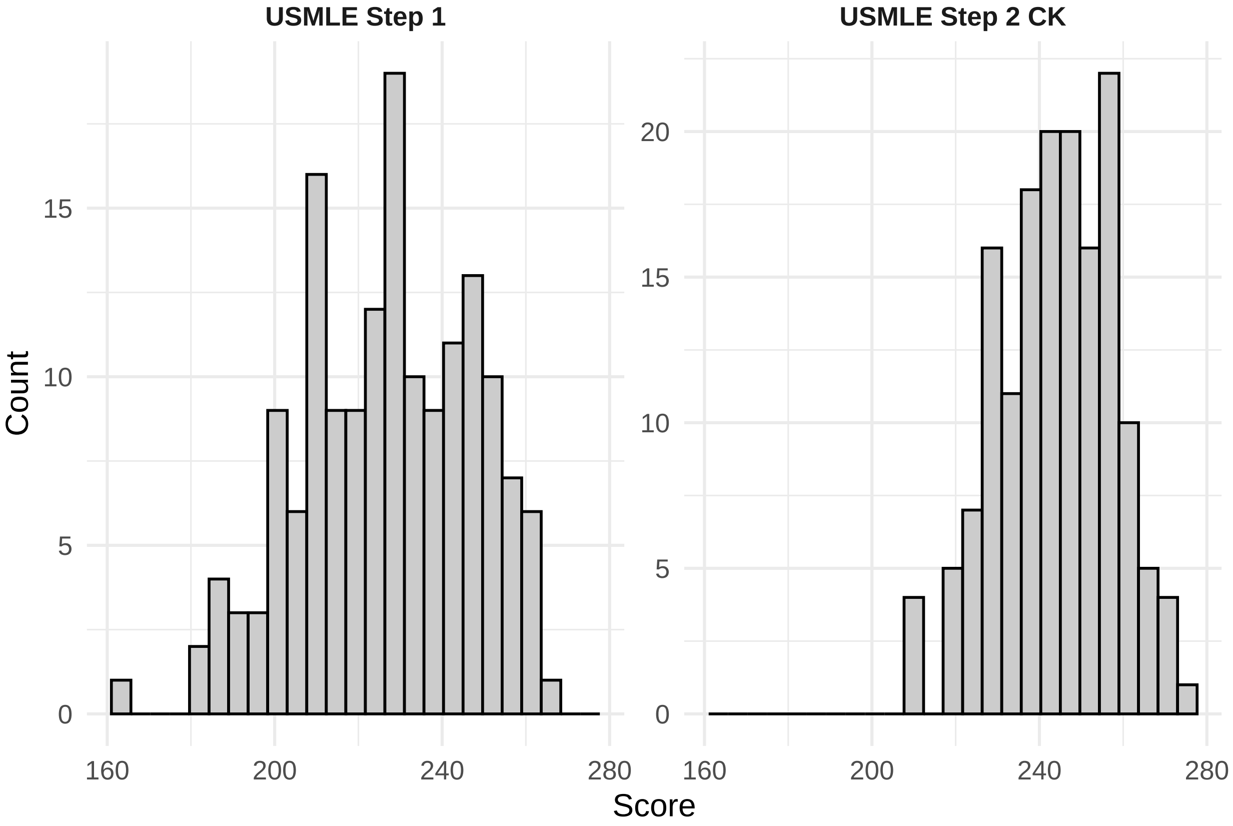

Figure 1: Distribution of USMLE Step 1 and Step 2 CK scores.

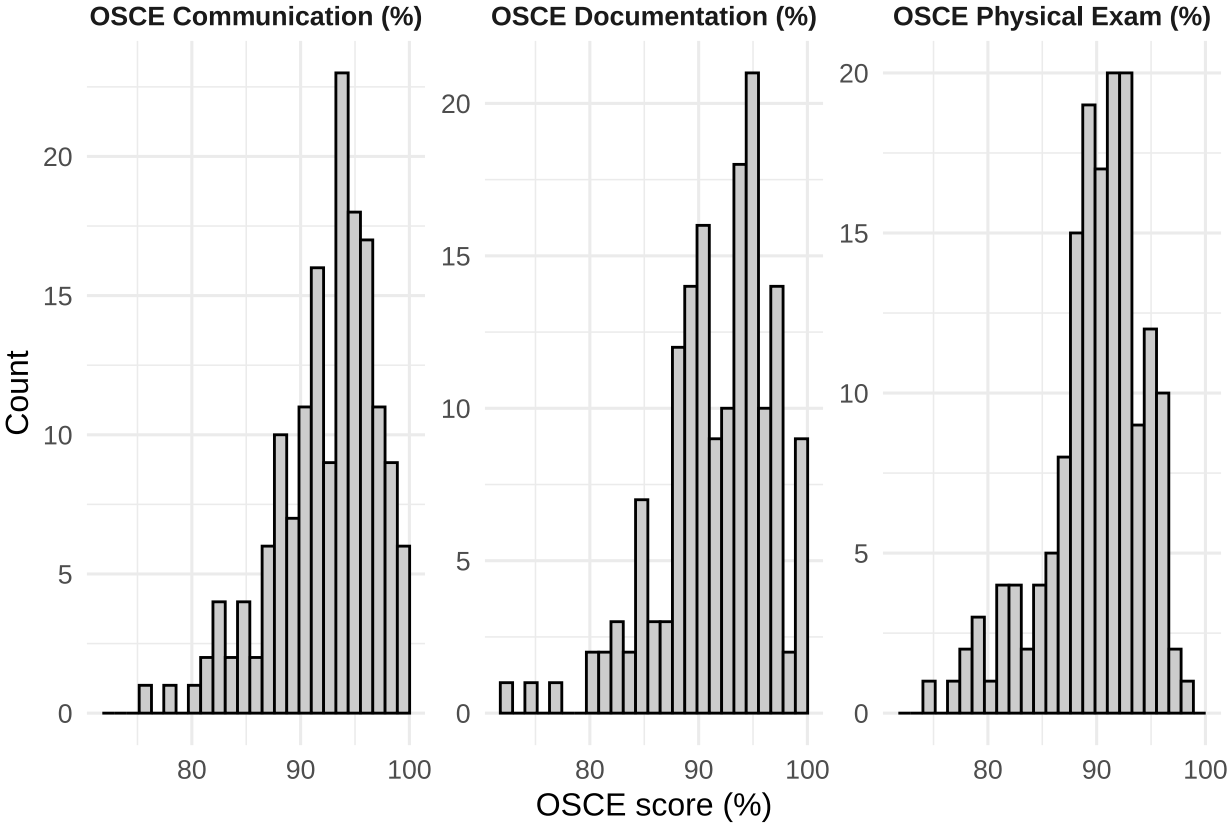

Figure 2: Distribution of OSCE domain scores (Communication, Documentation, Physical Exam).

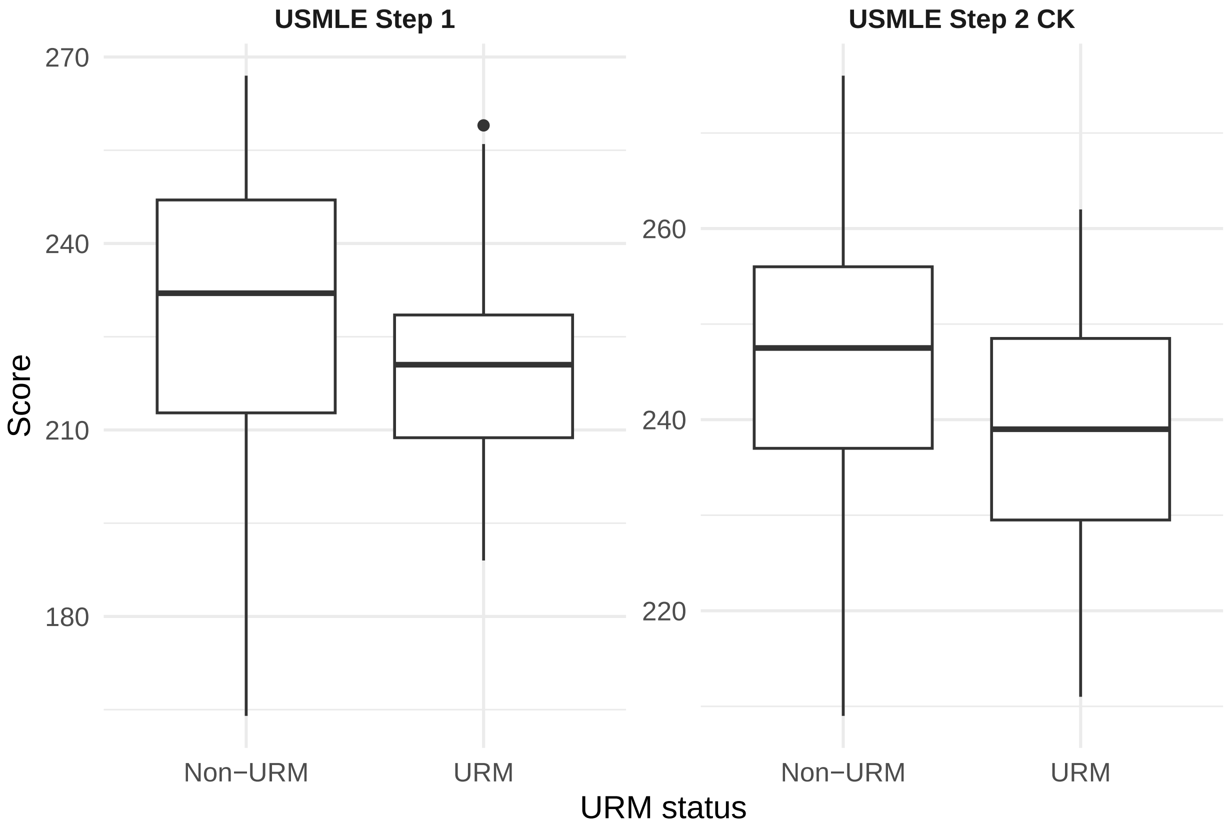

Figure 3: Boxplots of USMLE Step scores by URM status.

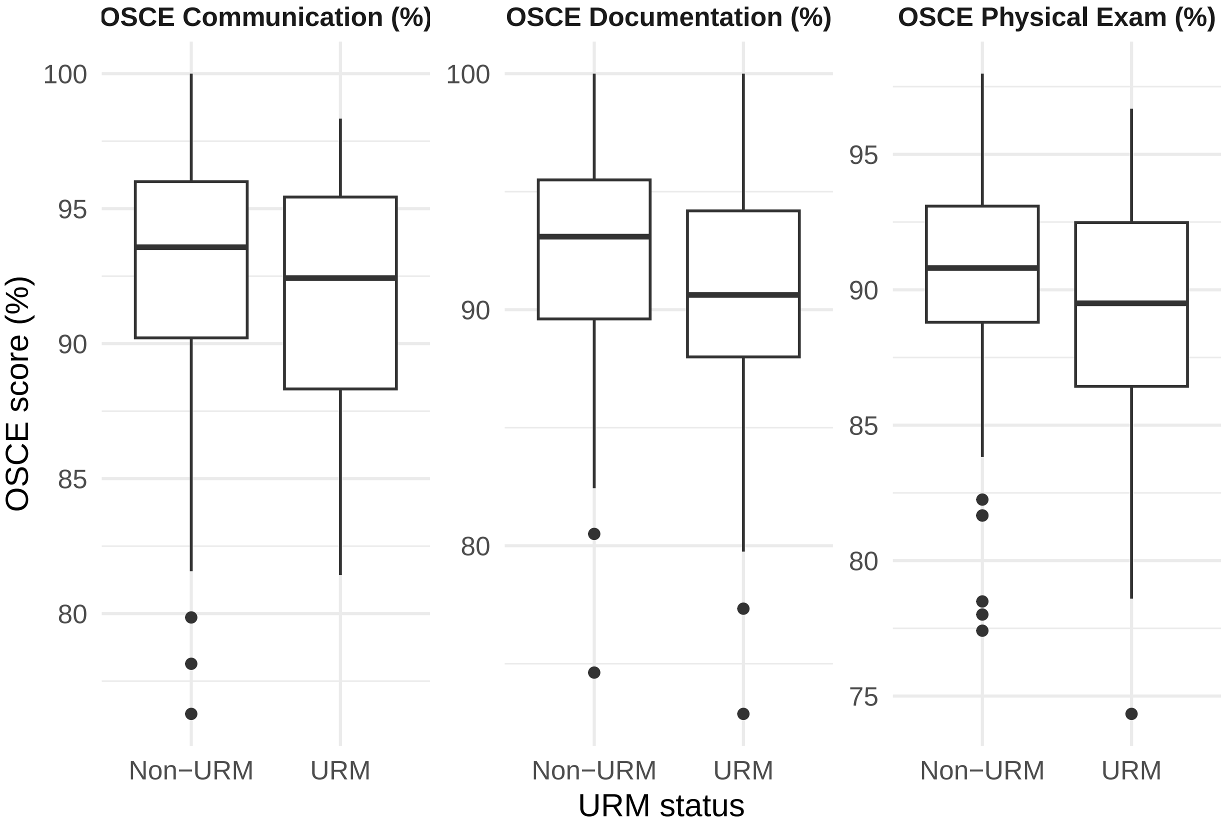

Figure 4: Boxplots of OSCE domain scores by URM status.

Code Appendix

```
## ---- Global chunk options & packages ---------------------
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

library(dplyr)
library(stringr)
library(tidyr)
library(table1)
library(broom)
library(knitr)
library(kableExtra)
library(purrr)
library(ggplot2)
## ---- Data import ----------------------------------------
# Read the original dataset of 199 students
#df <- read.csv("C:\\Users\\mahim\\OneDrive - The University of Colorado Denver\\BIOS6621_consulting\\F

df <- readr::read_csv(here::here("DataRaw", "BIOS6621_Guth_data.csv"))

## ---- Derive URM binary (Non-URM vs URM) -----------------
# URM variable contains 0, 1, and 3.
# Here, we define a binary URM indicator:
#   URM_bin = "URM" if URM == 1
#   URM_bin = "Non-URM" otherwise (0, 3, NA)
df <- df %>%
  mutate(
    URM_bin = factor(
      if_else(URM == 1, 1, 0),
      levels = c(0, 1),
      labels = c("Non-URM", "URM")
    )
  )


## -------- Clean gender variable (it had one "D") ---------
# Recode non-"F"/"M" values to missing (NA).
df <- df %>%
  mutate(
    gender_clean = case_when(
      gender %in% c("F", "M") ~ gender,
      TRUE ~ NA_character_
    )
  )

## ---- Identify OSCE score columns ------------------------
# OSCE domains:
#   Communication  - columns ending in "CommScore"
#   Documentation  - columns ending in "MedDocScore"
#   Physical Exam  - columns ending in "PhysExamScore"
comm_vars <- names(df)[str_detect(names(df), "CommScore$")]
doc_vars  <- names(df)[str_detect(names(df), "MedDocScore$")]
pe_vars   <- names(df)[str_detect(names(df), "PhysExamScore$")]
```

```r
#length(comm_vars); length(doc_vars); length(pe_vars)  # just to check

## ---- Aggregate OSCE scores to student-level means -------
# Compute student-level mean for each OSCE domain
df <- df %>%
  rowwise() %>%
  mutate(
    osce_comm_mean = if (all(is.na(c_across(all_of(comm_vars))))) NA_real_
                      else mean(c_across(all_of(comm_vars)), na.rm = TRUE),
    osce_doc_mean  = if (all(is.na(c_across(all_of(doc_vars)))))  NA_real_
                      else mean(c_across(all_of(doc_vars)), na.rm = TRUE),
    osce_pe_mean   = if (all(is.na(c_across(all_of(pe_vars)))))   NA_real_
                      else mean(c_across(all_of(pe_vars)), na.rm = TRUE)
  ) %>%
  ungroup()

## ---- Flag students missing ALL OSCE domains -------------
# Create an indicator for students missing all three OSCE domain means.
df <- df %>%
  mutate(
    all_osce_missing = if_else(
      is.na(osce_comm_mean) & is.na(osce_doc_mean) & is.na(osce_pe_mean),
      TRUE, FALSE
    )
  )

## ---- Define analytic dataset ----------------------------
# Restrict to students with at least one OSCE domain score.
# Keep variables needed for descriptive and inferential analyses.
analytic <- df %>%
  filter(!all_osce_missing) %>%     # keep only students with at least 1 OSCE domain
  select(
    ResearchID,
    URM_bin,
    race_desc,
    gender_clean,
    STEP1_ExamScore,
    STEP2CK_ExamScore,
    osce_comm_mean,
    osce_doc_mean,
    osce_pe_mean
  )
# nrow(analytic)   # should be ~160


## ---- Clean race (empty string to NA) --------------------
analytic <- analytic %>%
  mutate(
    race_desc = na_if(race_desc, "")
  )

##--------------merge race categories-----------
analytic <- analytic %>%
  mutate(
    race_simple = case_when(
      race_desc %in% c("White") ~ "White",

      race_desc %in% c("Asian", "Japanese") ~ "Asian",

      race_desc %in% c("African",
                       "African American",
                       "Black or African American") ~ "Black / African American",

      race_desc %in% c("American Indian or Alaska Native",
                       "American Indian or Alaskan Native") ~ "American Indian / Alaska Native",

      race_desc %in% c("Afro-Caribbean", "Other") ~ "Other",

      TRUE ~ NA_character_
    ),
    race_simple = factor(
      race_simple,
      levels = c("White", "Asian", "Black / African American",
                 "American Indian / Alaska Native", "Other")
    )
  )
# Sample sizes
n_orig    <- nrow(df)
n_analytic <- nrow(analytic)
n_all_osce_missing <- sum(df$all_osce_missing)

pct_analytic <- 100 * n_analytic / n_orig
pct_all_osce_missing <- 100 * n_all_osce_missing / n_orig

# URM counts / percents
urm_counts <- analytic |>
  count(URM_bin, name = "n") |>
  mutate(pct = 100 * n / sum(n))

n_nonurm <- urm_counts$n[urm_counts$URM_bin == "Non-URM"]
n_urm    <- urm_counts$n[urm_counts$URM_bin == "URM"]
pct_nonurm <- urm_counts$pct[urm_counts$URM_bin == "Non-URM"]
pct_urm    <- urm_counts$pct[urm_counts$URM_bin == "URM"]

# Gender by URM
gender_tab <- analytic |>
  count(URM_bin, gender_clean, name = "n") |>
  group_by(URM_bin) |>
  mutate(pct = 100 * n / sum(n))

# race by URM
race_tab <- analytic %>%
  count(URM_bin, race_simple) %>%
  group_by(URM_bin) %>%
  mutate(pct = round(100 * n / sum(n), 1))
```

```
# Step score summaries by URM (for narrative around Table 1)
step_summary <- analytic |>
  group_by(URM_bin) |>
  summarise(
    mean_step1 = mean(STEP1_ExamScore, na.rm = TRUE),
    sd_step1   = sd(STEP1_ExamScore, na.rm = TRUE),
    med_step1  = median(STEP1_ExamScore, na.rm = TRUE),
    min_step1  = min(STEP1_ExamScore, na.rm = TRUE),
    max_step1  = max(STEP1_ExamScore, na.rm = TRUE),
    mean_step2 = mean(STEP2CK_ExamScore, na.rm = TRUE),
    sd_step2   = sd(STEP2CK_ExamScore, na.rm = TRUE),
    med_step2  = median(STEP2CK_ExamScore, na.rm = TRUE),
    min_step2  = min(STEP2CK_ExamScore, na.rm = TRUE),
    max_step2  = max(STEP2CK_ExamScore, na.rm = TRUE)
  )

# OSCE domain summaries by URM
osce_summary <- analytic |>
  group_by(URM_bin) |>
  summarise(
    mean_comm = mean(osce_comm_mean, na.rm = TRUE),
    sd_comm   = sd(osce_comm_mean, na.rm = TRUE),
    med_comm  = median(osce_comm_mean, na.rm = TRUE),
    min_comm  = min(osce_comm_mean, na.rm = TRUE),
    max_comm  = max(osce_comm_mean, na.rm = TRUE),
    mean_doc  = mean(osce_doc_mean, na.rm = TRUE),
    sd_doc    = sd(osce_doc_mean, na.rm = TRUE),
    med_doc   = median(osce_doc_mean, na.rm = TRUE),
    min_doc   = min(osce_doc_mean, na.rm = TRUE),
    max_doc   = max(osce_doc_mean, na.rm = TRUE),
    mean_pe   = mean(osce_pe_mean, na.rm = TRUE),
    sd_pe     = sd(osce_pe_mean, na.rm = TRUE),
    med_pe    = median(osce_pe_mean, na.rm = TRUE),
    min_pe    = min(osce_pe_mean, na.rm = TRUE),
    max_pe    = max(osce_pe_mean, na.rm = TRUE)
  )
# ---- Label variables for Table 1 ----
label(df$URM_bin) <- "URM Status"
label(analytic$gender_clean)     <- "Gender"
label(analytic$race_desc) <- "Race"
label(analytic$STEP1_ExamScore)  <- "USMLE Step 1"
label(analytic$STEP2CK_ExamScore)<- "USMLE Step 2 CK"
label(analytic$osce_comm_mean)   <- "OSCE Communication"
label(analytic$osce_doc_mean)    <- "OSCE Documentation"
label(analytic$osce_pe_mean)     <- "OSCE Physical Exam"

units(analytic$osce_comm_mean)   <- "%"
units(analytic$osce_doc_mean)    <- "%"
units(analytic$osce_pe_mean)     <- "%"


label(analytic$race_simple) <- "Race"
```

```r
# ---- Build Table 1 (stratified by URM) ----
table1 <- table1(
  ~ gender_clean
    + race_simple
    + STEP1_ExamScore
    + STEP2CK_ExamScore
    + osce_comm_mean
    + osce_doc_mean
    + osce_pe_mean
  | URM_bin,
  data = analytic,
 overall = F
)

# Print nicely formatted Table 1
table1 %>%
  kable(
  caption = "Baseline demographic characteristics and assessment outcomes
  (OSCE domain scores and USMLE Step scores) among medical students, stratified by URM status."

  ) %>%
  kable_styling(full_width = FALSE, position = "center")

# ---- T-tests: OSCE domains by URM (unadjusted) ----------
# Unadjusted two-sample t-tests for each OSCE domain
t_comm <- t.test(osce_comm_mean ~ URM_bin, data = analytic)
t_doc  <- t.test(osce_doc_mean  ~ URM_bin, data = analytic)
t_pe   <- t.test(osce_pe_mean   ~ URM_bin, data = analytic)

# Extract summary info into a tidy table
tt_table <- bind_rows(
  tidy(t_comm)  %>% mutate(domain = "Communication"),
  tidy(t_doc)   %>% mutate(domain = "Documentation"),
  tidy(t_pe)    %>% mutate(domain = "Physical Exam")
) %>%
  select(domain, estimate, conf.low, conf.high, p.value) %>%
  mutate(
    estimate = round(estimate, 2),
    conf.low = round(conf.low, 2),
    conf.high = round(conf.high, 2),
    p.value = ifelse(p.value < 0.001, "<0.001", sprintf("%.3f", p.value))
  )

# Add group means separately
means_table <- tibble(
  domain = c("Communication", "Documentation", "Physical Exam"),
  mean_nonurm = c(
    t_comm$estimate[1],
    t_doc$estimate[1],
    t_pe$estimate[1]
  ),
  mean_urm = c(
    t_comm$estimate[2],
    t_doc$estimate[2],
    t_pe$estimate[2]
  )
) %>%
  mutate(across(c(mean_nonurm, mean_urm), ~ round(., 2)))

# Join means and t-test outputs into final Table 2
final_ttest_table <- tt_table %>%
  left_join(means_table, by = "domain") %>%
  select(
    Domain = domain,
    `Mean Non-URM` = mean_nonurm,
    `Mean URM` = mean_urm,
    `Mean Difference` = estimate,
    `95% CI` = conf.low,
    ` ` = conf.high,
    `p-value` = p.value
  ) %>%
  mutate(`95% CI` = paste0(`95% CI`, " to ", ` `)) %>%
  select(-` `)

# Print nicely
final_ttest_table %>%
  kable(
    caption = "Two-sample t-tests comparing mean OSCE domain scores
    between URM and Non-URM students.",
    align = "lcccccc"
  ) %>%
  kable_styling(full_width = FALSE, position = "center")
## ---- Adjusted linear models: OSCE domains ---------------
# Fit linear regression models for each OSCE domain:
# outcome = OSCE domain mean; predictors = URM, gender, Step 1, Step 2 CK

# model for Communication
m_comm <- lm(osce_comm_mean ~ URM_bin + gender_clean + STEP1_ExamScore + STEP2CK_ExamScore,
             data = analytic)

# model for Documentation
m_doc <- lm(osce_doc_mean ~ URM_bin + gender_clean + STEP1_ExamScore + STEP2CK_ExamScore,
            data = analytic)

# model for Physical exam
m_pe <- lm(osce_pe_mean ~ URM_bin + gender_clean + STEP1_ExamScore + STEP2CK_ExamScore,
           data = analytic)


# Collect results into a data frame
reg_osce <- bind_rows(
  tidy(m_comm, conf.int = TRUE) %>% mutate(domain = "Communication"),
  tidy(m_doc,  conf.int = TRUE) %>% mutate(domain = "Documentation"),
  tidy(m_pe,   conf.int = TRUE) %>% mutate(domain = "Physical Exam")
) %>%
  mutate(
    estimate = round(estimate, 2),
    conf.low = round(conf.low, 2),
    conf.high = round(conf.high, 2),
    p.value = ifelse(p.value < 0.001, "<0.001", sprintf("%.3f", p.value)),
    term = recode(term,
      "(Intercept)" = "Intercept",
      "URM_binURM" = "URM (vs Non-URM)",
      "gender_cleanM" = "Male (vs Female)",
      "STEP1_ExamScore" = "USMLE Step 1",
      "STEP2CK_ExamScore" = "USMLE Step 2 CK"
    )
  ) %>%
  mutate(
    ci = paste0(conf.low, " to ", conf.high)
  ) %>%
  select(
    Domain = domain,
    Predictor = term,
    Estimate = estimate,
    `95% CI` = ci,
    `p-value` = p.value
  )

reg_osce %>%
  arrange(Domain, Predictor) %>%
  kable(
    caption = "Adjusted linear regression models for OSCE domain scores
    with URM status, gender, and USMLE scores as predictors.",
    align = "llccc",
  ) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  collapse_rows(columns = 1)
## ---- FDR correction for URM effects ---------------------
# Extract URM p-values from the three adjusted OSCE models
p_vec <- c(
  broom::tidy(m_comm) %>% filter(term == "URM_binURM") %>% pull(p.value),
  broom::tidy(m_doc)  %>% filter(term == "URM_binURM") %>% pull(p.value),
  broom::tidy(m_pe)   %>% filter(term == "URM_binURM") %>% pull(p.value)
)

# Apply FDR (Benjamini-Hochberg) correction
p_adj <- p.adjust(p_vec, method = "fdr")


#  Create FDR summary table (Table 4)
fdr_table <- data.frame(
  Domain = c("Communication", "Documentation", "Physical Exam"),
  Raw_p_value = p_vec,
  FDR_adjusted_p_value = p_adj
) %>%
  mutate(
    Raw_p_value = sprintf("%.3f", Raw_p_value),
    FDR_adjusted_p_value = sprintf("%.3f", FDR_adjusted_p_value)
  )

# Nicely formatted table
fdr_table %>%
  kable(
    caption = "False Discovery Rate (FDR)-adjusted p-values
    for URM effects across OSCE domains.",
    align = "lcc",
  ) %>%
  kable_styling(
    full_width = FALSE,
    position = "center",
    bootstrap_options = c("striped", "hover")
  )
## ---- USMLE t-tests by URM (H2 unadjusted) ---------------
t_step1 <- t.test(STEP1_ExamScore ~ URM_bin, data = analytic)
t_step2 <- t.test(STEP2CK_ExamScore ~ URM_bin, data = analytic)


tt_step <- bind_rows(
  tidy(t_step1) %>% mutate(outcome = "USMLE Step 1"),
  tidy(t_step2) %>% mutate(outcome = "USMLE Step 2 CK")
) %>%
  select(outcome, estimate, conf.low, conf.high, p.value) %>%
  mutate(
    estimate = round(estimate, 2),
    conf.low = round(conf.low, 2),
    conf.high = round(conf.high, 2),
    p.value = ifelse(p.value < 0.001, "<0.001", sprintf("%.3f", p.value))
  )

# Add group means for Non-URM and URM for each Step exam
means_step <- tibble(
  outcome = c("USMLE Step 1", "USMLE Step 2 CK"),
  mean_nonurm = c(t_step1$estimate[1], t_step2$estimate[1]),
  mean_urm    = c(t_step1$estimate[2], t_step2$estimate[2])
) %>%
  mutate(across(c(mean_nonurm, mean_urm), ~ round(., 1)))

tt_step_table <- tt_step %>%
  left_join(means_step, by = "outcome") %>%
  mutate(
    ci = paste0(conf.low, " to ", conf.high)
  ) %>%
  select(
    Outcome = outcome,
    `Mean Non-URM` = mean_nonurm,
    `Mean URM` = mean_urm,
    `Mean Difference (Non-URM - URM)` = estimate,
    `95% CI` = ci,
    `p-value` = p.value
  )
```

```
tt_step_table %>%
  kable(
    caption = "Two-sample t-tests comparing USMLE Step scores
    between Non-URM and URM students.",
    align = "lccccc",
  ) %>%
  kable_styling(full_width = FALSE, position = "center")

## ---- USMLE models adjusted for gender (H2 adjusted) -----
m_step1 <- lm(STEP1_ExamScore ~ URM_bin + gender_clean, data = analytic)
m_step2 <- lm(STEP2CK_ExamScore ~ URM_bin + gender_clean, data = analytic)

urm_step <- bind_rows(
  tidy(m_step1, conf.int = TRUE) %>%
    filter(term == "URM_binURM") %>%
    mutate(outcome = "USMLE Step 1"),
  tidy(m_step2, conf.int = TRUE) %>%
    filter(term == "URM_binURM") %>%
    mutate(outcome = "USMLE Step 2 CK")
) %>%
  mutate(
    estimate = round(estimate, 1),
    conf.low = round(conf.low, 1),
    conf.high = round(conf.high, 1),
    p.value = ifelse(p.value < 0.001, "<0.001", sprintf("%.3f", p.value)),
    ci = paste0(conf.low, " to ", conf.high)
  ) %>%
  select(
    Outcome = outcome,
    `Adjusted Difference (URM - Non-URM)` = estimate,
    `95% CI` = ci,
    `p-value` = p.value
  )

urm_step %>%
  kable(
    caption = "Adjusted differences in USMLE Step scores
    between URM and Non-URM students.",
    align = "lccc"
  ) %>%
  kable_styling(full_width = FALSE, position = "center")
## ---- Correlations: OSCE domains vs USMLE (H3) -----------
# Define OSCE and USMLE variables and pretty labels
osce_vars <- c("osce_comm_mean", "osce_doc_mean", "osce_pe_mean")
osce_labels <- c(
  osce_comm_mean = "OSCE Communication",
  osce_doc_mean  = "OSCE Documentation",
  osce_pe_mean   = "OSCE Physical Exam"
)

step_vars <- c("STEP1_ExamScore", "STEP2CK_ExamScore")
step_labels <- c(
  STEP1_ExamScore   = "USMLE Step 1",
  STEP2CK_ExamScore = "USMLE Step 2 CK"
)

# Function to run cor.test for one OSCE × USMLE pair and extract stats
get_cor_row <- function(osce_var, step_var) {
  x <- analytic[[osce_var]]
  y <- analytic[[step_var]]
  ok <- complete.cases(x, y)

  ct <- cor.test(x[ok], y[ok])

  tibble(
    `OSCE Domain` = osce_labels[osce_var],
    `USMLE Exam`  = step_labels[step_var],
    r             = round(unname(ct$estimate), 2),
    `95% CI`      = sprintf("[%.2f, %.2f]", ct$conf.int[1], ct$conf.int[2]),
    `p-value`     = ifelse(ct$p.value < 0.001, "<0.001", sprintf("%.3f", ct$p.value))
  )
}

# Build correlation table for all OSCE × USMLE pairs
cor_osce_usmle <- cross_df(list(osce_var = osce_vars, step_var = step_vars)) %>%
  pmap_dfr(~ get_cor_row(..1, ..2))

# Nicely formatted table
cor_osce_usmle %>%
  kable(
    caption = "Pearson correlations between OSCE domain scores and USMLE Step scores.",
    align = c("l","l","c","c","c")
  ) %>%
  kable_styling(full_width = FALSE, position = "center")

## ---- Simple linear regressions: OSCE - USMLE associations (H3) -------
# Helper function: fit a simple linear model (USMLE outcome ~ OSCE predictor)
# and extract the slope, CI, and p-value
fit_model <- function(osce_var, step_var) {
  mod <- lm(reformulate(osce_var, response = step_var), data = analytic)
  tidied <- tidy(mod, conf.int = TRUE) %>%
    filter(term == osce_var) %>%
    mutate(
      Outcome = step_labels[[step_var]],
      Predictor = osce_labels[[osce_var]],
      Estimate = round(estimate, 3),
      CI = paste0("[", round(conf.low, 3), ", ", round(conf.high, 3), "]"),
      `p-value` = sprintf("%.3f", p.value)
    ) %>%
    select(Outcome, Predictor, Estimate, CI, `p-value`)
  return(tidied)
}

# apply to all combinations
reg_table <- cross_df(list(osce_var = osce_vars, step_var = step_vars)) %>%
  pmap_dfr(~ fit_model(..1, ..2))
```

```r
# nicely formatted table
reg_table %>%
  kable(
    caption = "Associations between OSCE domain scores and USMLE Step scores from simple linear regress
    align = c("l",  "l",  "c",  "c",  "c")
  ) %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  collapse_rows(columns = 1)
## ---- Histograms: USMLE Step scores ----------------------
# Long format for USMLE Step 1 and Step 2 CK histograms.
usmle_long <- analytic %>%
  select(`USMLE Step 1` = STEP1_ExamScore,
         `USMLE Step 2 CK` = STEP2CK_ExamScore) %>%
  pivot_longer(cols = everything(),
               names_to = "Exam",
               values_to = "Score") %>%
  filter(!is.na(Score))

# plot
ggplot(usmle_long, aes(x = Score)) +
  geom_histogram(bins = 25, color = "black", fill = "grey80") +
  facet_wrap(~ Exam, scales = "free_y") +
  labs(x = "Score", y = "Count") +
  theme_minimal(base_size = 12) +
  theme(
    strip.text = element_text(face = "bold"),
    plot.caption = element_text(hjust = 0)
  )
## ---- Histograms: OSCE domain scores ---------------------
# Long format for OSCE histograms (communication, documentation, PE)
osce_long <- analytic %>%
  select(
    `OSCE Communication (%)` = osce_comm_mean,
    `OSCE Documentation (%)` = osce_doc_mean,
    `OSCE Physical Exam (%)` = osce_pe_mean
  ) %>%
  pivot_longer(cols = everything(),
               names_to = "Domain",
               values_to = "Score") %>%
  filter(!is.na(Score))

ggplot(osce_long, aes(x = Score)) +
  geom_histogram(bins = 25, color = "black", fill = "grey80") +
  facet_wrap(~ Domain, scales = "free_y") +
  labs(x = "OSCE score (%)", y = "Count") +
  theme_minimal(base_size = 12) +
  theme(
    strip.text = element_text(size = 10, face = "bold"),
    plot.caption = element_text(hjust = 0)
  )

## ---- Boxplots: USMLE by URM status ----------------------
# Long format for USMLE boxplots stratified by URM status
```

```r
usmle_by_urm <- analytic %>%
  select(
    URM_bin,
    `USMLE Step 1`   = STEP1_ExamScore,
    `USMLE Step 2 CK` = STEP2CK_ExamScore
  ) %>%
  pivot_longer(
    cols = c(`USMLE Step 1`, `USMLE Step 2 CK`),
    names_to = "Exam",
    values_to = "Score"
  ) %>%
  filter(!is.na(Score), !is.na(URM_bin))

# plot
ggplot(usmle_by_urm, aes(x = URM_bin, y = Score)) +
  geom_boxplot() +
  facet_wrap(~ Exam, scales = "free_y") +
  labs(x = "URM status", y = "Score") +
  theme_minimal(base_size = 12) +
  theme(
    strip.text   = element_text(face = "bold"),
    plot.caption = element_text(hjust = 0)
  )

## ---- Boxplots: OSCE domains by URM status ---------------
# Long format for OSCE boxplots stratified by URM status
osce_by_urm <- analytic %>%
  select(
    URM_bin,
    `OSCE Communication (%)` = osce_comm_mean,
    `OSCE Documentation (%)` = osce_doc_mean,
    `OSCE Physical Exam (%)` = osce_pe_mean
  ) %>%
  filter(!is.na(URM_bin)) %>%              # <- keep only defined URM
  pivot_longer(
    cols = -URM_bin,
    names_to = "Domain",
    values_to = "Score"
  ) %>%
  filter(!is.na(Score))


ggplot(osce_by_urm, aes(x = URM_bin, y = Score)) +
  geom_boxplot() +
  facet_wrap(~ Domain, scales = "free_y") +
  labs(x = "URM status", y = "OSCE score (%)") +
  theme_minimal(base_size = 12) +
  theme(
    strip.text   = element_text(size = 10, face = "bold"),
    plot.caption = element_text(hjust = 0)
  )
```