Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Station-Based Performance in Objective Structured Clinical Examinations: A Comparative Study in a large Ukrainian Cohort (1233 students)

Boyko Yulia
Bogomolets National Medical University
Kuchyn Yurii
Bogomolets National Medical University
Korotia Mykola
Bogomolets National Medical University
Melnyk Veronika
Bogomolets National Medical University
Sereda Sergiy
Bogomolets National Medical University
Yyval Mykola
vyval.mykola@gmail.com

SI «Scientific-Practical Center of Endovascular Neuroradiology of the NAMS of Ukraine»

Research Article

Keywords: medical education, objective structured clinical examination (OSCE), station-level performance

Posted Date: June 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9669534/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Objective Structured Clinical Examinations (OSCEs) are widely used to assess clinical competence in medical education; however, detailed analysis of station-level performance and comparability across specialties remains limited. This study aimed to evaluate station-level performance within specialties during OSCE-1 and to compare overall examination outcomes across different educational programmes.

Methods

This retrospective study included 1233 students from four programmes: Medicine (n=989), Medicine for the Armed Forces (n=40), Pediatrics (n=137), and Physical Therapy and Rehabilitation (n=67). Within each specialty, descriptive statistics, Wilcoxon signed-rank tests with Benjamini–Hochberg correction, and Spearman correlation analysis were applied. Between-group comparisons were performed using the Mann–Whitney U test with adjustment for multiple testing.

Results

Across all programmes, higher performance was observed in procedural and emergency stations, while lower scores and greater variability were consistently found in electrocardiogram interpretation. The proportion of students scoring below 60 points was highest in these domains, reaching over 10% in several cohorts. Pairwise comparisons confirmed significant differences between stations within specialties, whereas correlation analysis indicated predominantly weak to moderate associations, suggesting multidimensional competency assessment. Between-group analysis demonstrated broadly comparable overall results, with Medicine outperforming Pediatrics and Physical Therapy and Rehabilitation ($p<0.05$), while other comparisons were not statistically significant.

Conclusions

OSCE-1 provides detailed differentiation of competency domains at the station level while maintaining overall comparability across specialties through a unified scoring system. Combined analysis of station-specific and integrated outcomes offers valuable insights for targeted curriculum development and assessment design.

Clinical trial number: not applicable.

Introduction

The objective structured clinical examination (OSCE) has become a cornerstone of contemporary medical education, representing a paradigm shift from predominantly theoretical knowledge-based assessment toward the direct evaluation of clinical competence in standardized, performance-based environments (1, 2). Since its introduction by Harden and colleagues in the 1970s (3), the OSCE has been

widely adopted across undergraduate and postgraduate medical training globally (4-6), offering a structured format in which learners rotate through a series of stations designed to assess discrete clinical tasks under controlled conditions. This approach enables the systematic evaluation of a broad range of competencies, including history taking, physical examination, clinical reasoning, procedural skills, and communication, with scoring based on predefined checklists to enhance objectivity and reproducibility (7). Over the past decades, OSCEs have been increasingly integrated into high-stakes licensing examinations and institutional assessments worldwide (6, 8-11), reflecting a broader shift toward competency-based medical education and the need for reliable tools that capture real-world clinical performance beyond theoretical knowledge (12).

Ukraine has introduced the OSCE as a key component of the examination, complementing traditional written assessments (13, 14). The implementation of the OSCE reflects an effort to align national medical education standards with international practices and to ensure that graduates demonstrate required applied clinical competencies. Ukrainian medical universities have progressively adopted OSCE-based formats across multiple specialties, including medicine, pediatrics, dentistry, and rehabilitation medicine, often utilizing simulation technologies and standardized clinical scenarios to approximate real-world practice (14-17).

Despite this rapid adoption, the evidence base describing the structure, outcomes, and educational impact of OSCE implementation in Ukraine remains comparatively limited and fragmented. Existing publications mainly focus on descriptive accounts of implementation, specific institutional experiences, or methodological considerations in selected specialties, rather than providing comprehensive comparative analyses across the OSCE stations (14-18). The predominance of aggregated scoring without detailed analysis limits the ability to identify specific areas of strength and weakness within clinical training, thereby constraining evidence-based curriculum development. Notably, the potential use of detailed station-level data to inform targeted educational improvements remains underexplored.

In the Ukrainian medical education system, undergraduate training in Medicine, Medicine for the Armed Forces, and Pediatrics is structured as a six-year programme. Students undertake OSCE-1 and OSCE-2 at the end of third and sixth years, respectively. OSCE-1 and OSCE-2 results contribute to the final ranking used for specialty training allocation upon completion of medical school. At the same time, the Physical Therapy and Rehabilitation programme is organized as a four-year undergraduate course (OSCE-1) followed by a two-year master's course (OSCE-2).

This study aimed to evaluate OSCE-1 station-level performance within each specialty (primary endpoint) and to compare overall examination performance across specialties as a secondary endpoint.

Methods

Participants

This retrospective study was based on publicly available data extracted from the institutional registry of Bogomolets National Medical University (Kyiv, Ukraine). The study was conducted in strict accordance with the Declaration of Helsinki and national ethics and data protection regulations. Ethical approval was granted by the Ethics Committee of Bogomolets National Medical University (approval № 205, April 27, 2026). The ethics committee waived the requirement of written informed consent for participation due to the retrospective nature of the study. All participants had previously provided informed consent for the use of their educational and assessment data for research and quality assurance purposes in accordance with institutional policies.

The dataset included all students who undertook the OSCE-1 exam in 2025, comprising a total of 1233 individuals. Among them, 989 were enrolled in the Medicine programme, 40 in Medicine for the Armed Forces, 137 in Pediatrics, and 67 in Physical Therapy and Rehabilitation. No additional exclusion criteria were applied. Sample size was determined by the number of eligible subjects.

OSCE-1 exam

In all cases, the examination consisted of multiple stations designed to assess distinct domains of practical clinical competence. The full list of stations is provided in **Table 1**. Each station was scored on a scale from 0 to 100. For each candidate, the mean of station scores was calculated and subsequently multiplied by two to obtain the final examination score on a 200-point scale. A final score below 60% (i.e., <120 points) was considered a failing result and required resitting of the examination as a prerequisite for progression to the subsequent stage of training.

Station-level performance (within-group comparison)

Within each specialty, station-level performance was compared to assess differences in student achievement across individual stations. Descriptive statistics was calculated to characterize the distribution of results, including calculation of the mean, standard deviation (SD), standard error (SE), confidence intervals (CI), median, minimum and maximum values, and interquartile range (IQR), as well as visualizing results via violin plots. Additionally, we calculated the number of students who scored below 60 points for each of the stations. This was performed to identify areas requiring improvement in practical skills and to reduce the likelihood of retaking the entire examination.

Given that the distribution of the analyzed variables deviated from normality, non-parametric statistical methods were applied. Pairwise comparisons between all possible station combinations were performed using the Wilcoxon signed-rank test (19). To account for multiple testing, p-values were adjusted using the Benjamini–Hochberg procedure (20).

Spearman correlation analysis (21) was conducted to evaluate associations between performance across different stations. This approach allowed assessment of whether higher performance at one station was accompanied by higher or lower performance at other stations. Such relationships were

interpreted as statistical associations only, without inference of causality, and were used to explore whether stations reflected shared underlying competencies or assessed independent skill domains.

Between-group comparison

As an explorative analysis, between-group comparisons were performed across four specialties (Medicine, Medicine for the Armed Forces, Pediatrics, and Physical Therapy and Rehabilitation). To ensure comparability, the analysis was restricted to the final examination score expressed on the standardized 200-point scale, which was treated as an integrated measure of overall performance. This approach allows valid comparison between groups with differing numbers and compositions of stations, as the final score reflects a unified assessment framework defined by institutional criteria. Importantly, within the Ukrainian system, OSCE results expressed on the 200-point scale contribute to a single unified ranking for specialty training allocation, in which graduates of General Medicine, Medicine for the Armed Forces, and Pediatrics are ranked together.

Given the non-normal distribution of the data and the marked imbalance in group sizes, non-parametric methods were applied. Specifically, pairwise comparisons between specialties were conducted using the Mann–Whitney U test (22) followed by the Benjamini–Hochberg procedure (20).

Results

Station-level performance (within-group comparison)

A total of 1233 students were included in the analysis. Station-based performance is illustrated in **Figure 1**, with detailed descriptive statistics provided in **Supplementary Tables 1–4**. Between-station comparisons and correlation analysis results are presented in **Supplementary tables 5-8** and **Supplementary figures 1-4**, respectively.

Medicine

Within the Medicine programme, overall performance was high across most stations, with mean scores exceeding 90 points in several domains, including M8 Trauma care (94.49 points), M9 Blood group assessment (95.23), and M10 Basic procedural skills (94.78). In contrast, lower performance was observed in M5 Electrocardiogram interpretation (77.73), M2 Anamnesis taking (therapy) (80.10), and M3 Physical examination in therapy (82.87). The widest variability was also noted for M5 Electrocardiogram interpretation (SD 22.49), indicating substantial heterogeneity in student performance.

A focused analysis of students scoring below 60 points at individual stations revealed distinct patterns of underperformance across programmes (**Table 2**). In the Medicine programme, failure rates were generally low across most stations; however, a marked exception was observed for electrocardiogram interpretation (M5), where 11.43% of students scored below the threshold, representing the highest failure rate among all assessed stations. Other stations with comparatively elevated proportions of low

scores included physical examination in therapy (M3, 3.84%), blood pressure measurement (M4, 3.24%), and anamnesis taking in therapy (M2, 2.73%). In contrast, several stations demonstrated near-complete proficiency, including trauma care (M8) and basic procedural skills (M10), where no students scored below 60 points.

Pairwise comparisons demonstrated statistically significant differences between the majority of station combinations after Benjamini–Hochberg correction ($p<0.05$), with particularly large effect sizes observed when comparing lower-performing stations (e.g., M5, M2) to higher-performing procedural and emergency stations (e.g., M8–M10). Please see **Supplementary table 5**.

Spearman correlation analysis (**Supplementary figure 1**) demonstrated generally weak to moderate associations between station scores, suggesting that individual stations largely assess distinct domains of competence rather than a single unified skill construct.

Medicine for the Armed Forces

For Medicine for the Armed Forces, the highest scores were observed in surgery-related stations, including M6 (anamnesis taking) and M7 (physical examination). In contrast, lower scores were noted in M2 (anamnesis taking) and M5 (electrocardiogram interpretation), indicating comparatively weaker performance.

Analysis of scores below 60 points (**Table 2**) revealed that failures were largely confined to a small subset of stations. The highest proportions were observed in M2 and M3, each affecting 7.5% of students, while isolated cases were seen in M1, M4, and M5. Notably, no failures were recorded in surgical, emergency, or procedural stations (M6–M12), suggesting that these competencies were reliably achieved across the cohort.

Pairwise comparisons confirmed substantial differences between stations, particularly between lower-performing domains and higher-performing tasks ($p<0.05$). Detailed results are presented in **Supplementary table 6**.

Spearman correlation analysis (**Supplementary figure 2**) showed predominantly weak associations between stations.

Pediatrics

The Pediatrics cohort demonstrated that the highest scores were observed in foundational pediatric tasks such as P1 (child anthropometry), where results were consistently near the upper limit, as well as in P12 (cardiopulmonary resuscitation). In contrast, P5 (electrocardiogram interpretation) represented the lowest-performing station, accompanied by substantial variability.

The distribution of low scores further emphasized the aforementioned pattern (please see **Table 2**). Electrocardiogram interpretation emerged as the primary point of difficulty, with 10.95% of students

scoring below 60 points. Additional challenges were observed in P4 (blood pressure measurement), while most other stations showed only sporadic failures. Importantly, several core pediatric competencies, including child anthropometry, infant feeding, and child physical examination, showed no instances of low performance.

Pairwise comparisons revealed widespread statistically significant differences between stations ($p<0.05$). These findings are detailed in **Supplementary table 7**.

Correlation analysis (**Supplementary figure 3**) demonstrated generally weak relationships between station scores, indicating that performance in pediatric OSCE stations reflects multiple relatively independent competencies.

Physical Therapy and Rehabilitation

In contrast to the other programmes, the Physical Therapy and Rehabilitation cohort exhibited a markedly uniform performance profile. Mean scores across all stations were relatively close, without clear outliers in either direction, suggesting a balanced level of performance across assessed domains. No station demonstrated either exceptionally high or distinctly low mean values relative to the rest.

However, analysis of low-performing students (**Table 2**) revealed that several stations showed comparable proportions of scores below 60 points, including PTR1, PTR3, PTR4, and PTR6, each affecting a noticeable subset of students. Unlike the other programmes, there were no stations entirely free of low scores, indicating that underperformance was not confined to a single domain but rather distributed across competencies.

Consistent with this observation, pairwise comparisons did not identify any statistically significant differences between stations after correction for multiple testing ($p>0.05$), reinforcing the absence of strong performance gradients within the cohort (**Supplementary table 8**).

Spearman correlation analysis (**Supplementary figure 4**) demonstrated weak to moderate associations between stations, suggesting partial overlap between functional skills while still supporting the presence of distinct competency components.

Between-group comparison

Final examination scores differed across programmes, reflecting variation in overall performance when assessed using the unified 200-point scale (**Figure 2, Supplementary table 9**).

Students in the Medicine programme achieved the highest results, with a mean score of 178.65, followed closely by Medicine for the Armed Forces (176.43). Lower overall performance was observed in Pediatrics (171.55) and Physical Therapy and Rehabilitation (168.94), which also demonstrated greater dispersion of results.

Pairwise comparisons revealed that the Medicine cohort significantly outperformed both Pediatrics ($p<0.05$) and Physical Therapy and Rehabilitation ($p<0.05$), with moderate effect sizes ($\eta^2 \approx 0.13–0.14$). In contrast, no statistically significant difference was observed between Medicine and Medicine for the Armed Forces ($p>0.05$), indicating comparable overall performance between these closely related training pathways. There were no significant differences in other comparisons as well. Please refer to **Supplementary table 10**.

Discussion

Overall, the present study provided a comprehensive analysis of OSCE-1 performance across multiple medical and allied health programmes in the leading Ukrainian medical university, revealing consistent patterns at both station and programme levels. The findings demonstrated that students perform reliably well in structured procedural and emergency tasks, while comparatively lower performance is observed in cognitively demanding domains, particularly electrocardiogram interpretation and components of clinical reasoning such as anamnesis taking and physical examination in therapy-oriented contexts. This pattern was consistently reproduced across programmes, indicating that the observed differences are unlikely to be programme-specific and instead reflect broader characteristics of training and assessment.

One of the most prominent findings is the persistent difficulty associated with electrocardiogram interpretation, which demonstrated both the lowest mean scores and the highest proportion of students scoring below the threshold across multiple cohorts. It suggests that this competency may be insufficiently developed at the stage corresponding to OSCE-1. In contrast, emergency and procedural stations, including trauma care and cardiopulmonary resuscitation, consistently demonstrated high performance and near absence of failure cases, indicating that these skills are effectively taught and retained. Taken together, these results point toward an imbalance in training emphasis, where algorithmic and hands-on competencies may be prioritized over interpretative and integrative clinical skills.

From a structural perspective, the weak to moderate correlations between stations further support the interpretation that OSCE-1 assesses a multidimensional construct of clinical competence rather than a single unified ability. This has important implications for both assessment design and curriculum development, as it suggests that deficiencies in one domain cannot be compensated by strengths in another, reinforcing the need for targeted educational interventions.

For the between-programme analysis, overall examination outcomes across programmes were broadly comparable, with only a limited number of statistically significant differences observed. Despite differences in curriculum structure and station composition, these findings should be interpreted in the context of the Ukrainian medical educational system, where final OSCE results, expressed on a unified 200-point scale, are used as a common metric for ranking graduates across specialties for subsequent training allocation.

The findings of this study have direct implications for medical education in Ukraine. First, the consistent underperformance in electrocardiogram interpretation suggests a need to reconsider how this competency is taught, potentially shifting toward more practical and case-based approaches rather than isolated theoretical instruction. Second, the observed variability in therapy-oriented communication and examination skills indicates that greater emphasis should be placed on early clinical reasoning and patient interaction, particularly in non-procedural contexts. Third, the strong performance in emergency and procedural stations provides a positive indication that current training approaches in these domains are effective and may serve as a model for strengthening other areas of the curriculum.

Several limitations of the present study should be acknowledged. The study is based on data from a single institution, which may limit generalizability to other educational settings. The retrospective design precludes control over potential confounding variables, including differences in teaching methods, examiner variability, and prior student experience. Additionally, while the use of a standardized 200-point scale enables comparison between programmes, underlying differences in station composition may still influence the results. Finally, the analysis is limited to quantitative performance metrics and does not capture qualitative aspects of student competence or examiner judgment.

Future research should aim to expand this analysis to multiple institutions to assess the reproducibility of the observed patterns across different educational environments. Longitudinal studies following student performance across multiple stages of training would be particularly valuable in determining whether the identified gaps persist or are resolved over time. Furthermore, integrating qualitative assessment methods, including structured examiner feedback and student self-assessment, may provide deeper insight into the underlying causes of underperformance in specific domains. Finally, targeted educational interventions addressing identified weaknesses, particularly in electrocardiogram interpretation and clinical reasoning, should be developed and prospectively evaluated to determine their effectiveness.

Conclusions

OSCE-1 revealed clear differences in performance across individual stations within each specialty, confirming that the examination captures heterogeneous components of clinical competence rather than a single unified skill. This allows identification of specific domains requiring improvement at the level of training programmes.

At the same time, overall examination scores were largely comparable across specialties, with only limited differences reaching statistical significance. Within the framework of the Ukrainian system, where final OSCE results are used as a unified metric for graduate ranking, the 200-point scale appears to provide a functionally consistent basis for comparison despite differences in curriculum and station composition.

These findings support the use of combined station-level and overall score analysis as complementary approaches for evaluating both detailed competency profiles and integrated educational outcomes.

Abbreviations

OSCE, Objective Structured Clinical Examination

Declarations

Ethics approval and consent to participate

Ethical approval was granted by the Ethics Committee of Bogomolets National Medical University (approval № 205, April 27, 2026). The ethics committee waived the requirement of written informed consent for participation due to the retrospective nature of the study.

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

BY, KY, KM, MV, and SS are employed at Bogomolets National Medical University, Kyiv, Ukraine.

Funding

None

Authors' contributions

Conceptualization: BY, KY, KM, MV, SS, VM; methodology: BY, KY, KM, MV, SS, VM; formal analysis: BY, SS, VM; resources: BY, KY, KM, MV, SS, VM; data writing – original draft preparation: BY, VM; writing – review and editing: BY, KY, KM, MV, SS, VM; supervision: KY, VM; project administration: BY, KY, KM, MV, SS, VM. All authors read and approved the final manuscript.

Acknowledgements

The authors used Microsoft Copilot for grammar and style checking. The authors reviewed all suggestions and take full responsibility for the final version of the manuscript.

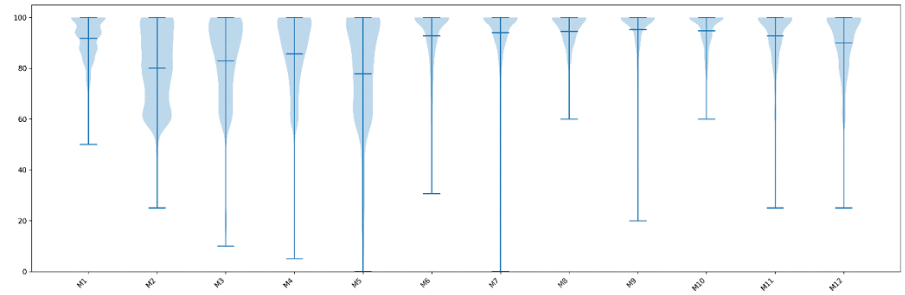
References

1. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. Med Teach. 2013;35(9):e1437-46.
2. Al-Hashimi K, Said UN, Khan TN. Formative Objective Structured Clinical Examinations (OSCEs) as an Assessment Tool in UK Undergraduate Medical Education: A Review of Its Utility. Cureus. 2023;15(5):e38519.
3. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1(5955):447-51.
4. Skrzypek A, Szeliga M, Stalmach-Przygoda A, Górski S, Kowalska B, Kocurek A, et al. The Objective Structured Clinical Examination (OSCE) from the perspective of 3rd year's medical students—a pilot study. Folia Medica Cracoviensia. 2017.
5. Yazbeck Karam V, Park YS, Tekian A, Youssef N. Evaluating the validity evidence of an OSCE: results from a new medical school. BMC Med Educ. 2018;18(1):313.
6. Henderson C, Stirling K, Jones A, Smith LH, Daffada P, Thomas J, et al. Evaluating OSCE assessments in UK advanced practice curricula: insights from educators. Br J Nurs. 2026;35(2):104-12.
7. Hijazi M, Downing SM. Objective structured clinical examinations as an assessment method in residency training: practical considerations. Ann Saudi Med. 2008;28(3):192-9.
8. Müller S, Dahmen U, Settmacher U. [Application of the Objective Structured Clinical Examination (OSCE) in German Medical Schools: An Inventory]. Gesundheitswesen. 2018;80(12):1099-103.
9. G P, Hareesh GSR, Jasti LM, Chandrika H. A Comparative Analysis of Objective Structured Clinical Examinations (OSCEs) and Traditional Written Exams in Medical Education for Assessing Clinical Competence. Cureus. 2026;18(2):e102867.
10. Rangiah S. The objective structured clinical examination in family medicine training: A comprehensive review for the South African context. S Afr Fam Pract (2004). 2026;68(2):e1-e6.
11. Onwudiegwu U. OSCE: DESIGN, DEVELOPMENT AND DEPLOYMENT. J West Afr Coll Surg. 2018;8(1):1-22.
12. Gupta DK, Chaudhuri A. Reframing Assessment in Medical Education: Toward Emotionally Sustainable Assessment in the Era of Competency-based Medical Education. Current Medical Issues. 2026;24(2):281-6.
13. Marichereda V, Melnychenko M, Buryachkovskyi E, Kashtalian M, Eliy L, Khatsko V. Objective structured clinical examination as a component of state certification of resident doctors specializing in pediatric surgery, based on the example of Odessa National Medical University. The Ukrainian Journal of Clinical Surgery. 2026;93(1):102-7.
14. Troyanovska O. The Current Status of Medical Education in Ukraine. Igaku Kyoiku/Medical Education (Japan). 2022;53(6):521-4.
15. Marynchak O. Objective Structured Clinical Examination as Adequate MEthods For Assessing The Professional Competence For Intership Doctor. Art of Medicine. 2020:116-20.

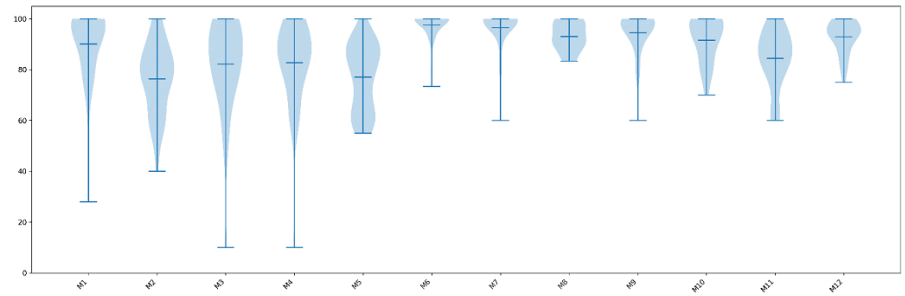
16. Remez D. Overview of Medical Education in Ukraine. Igaku Kyoiku/Medical Education (Japan). 2022;53(6):502-11.
17. Vitomskyi V, Klavina A, Mruga M, Molik B, Gavreliuk S, Reklaitiene D, et al. Physical therapy graduate students' and examiners' perception of objective structured clinical examination: a feedback for process improvement. Health, sport, rehabilitation. 2022;8(2):63-77.
18. SHOSTAKOVYCH-KORETSKA L, CHERHINETS A, LYTVYN K, MAVRUTENKOV V, HAYDUK T, CHYKARENKO Z, et al. Experience in developing simulation script for performance of Objective Structured Clinical Examination (OSCE). EUROPEAN HUMANITIES STUDIES: State and Society. 2019(1 (I)):178-92.
19. Wilcoxon F. Individual comparisons by ranking methods. Biometrics bulletin. 1945;1(6):80-3.
20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289-300.
21. Spearman C. The proof and measurement of association between two things. 1961.
22. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics. 1947:50-60.

Figures

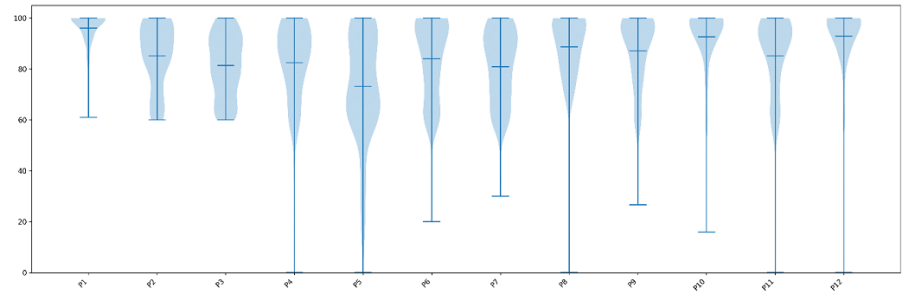
A. Medicine (n=989)



B. Medicine for the Armed Forces (n=40)



C. Pediatrics (n=137)



D. Physical Therapy and Rehabilitation (n=67)

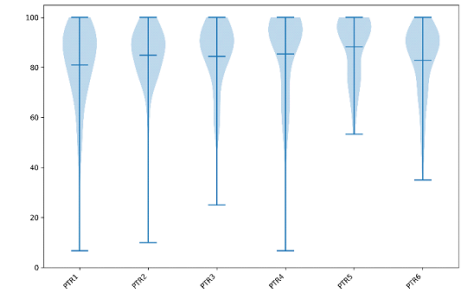


Figure 1

Station-based performance. (A) Medicine (n=989); (B) Medicine for the Armed Forces (n=40); (C) Pediatrics (n=137); (D) Physical Therapy and Rehabilitation (n=67). Please refer to **Table 1** for the station details.

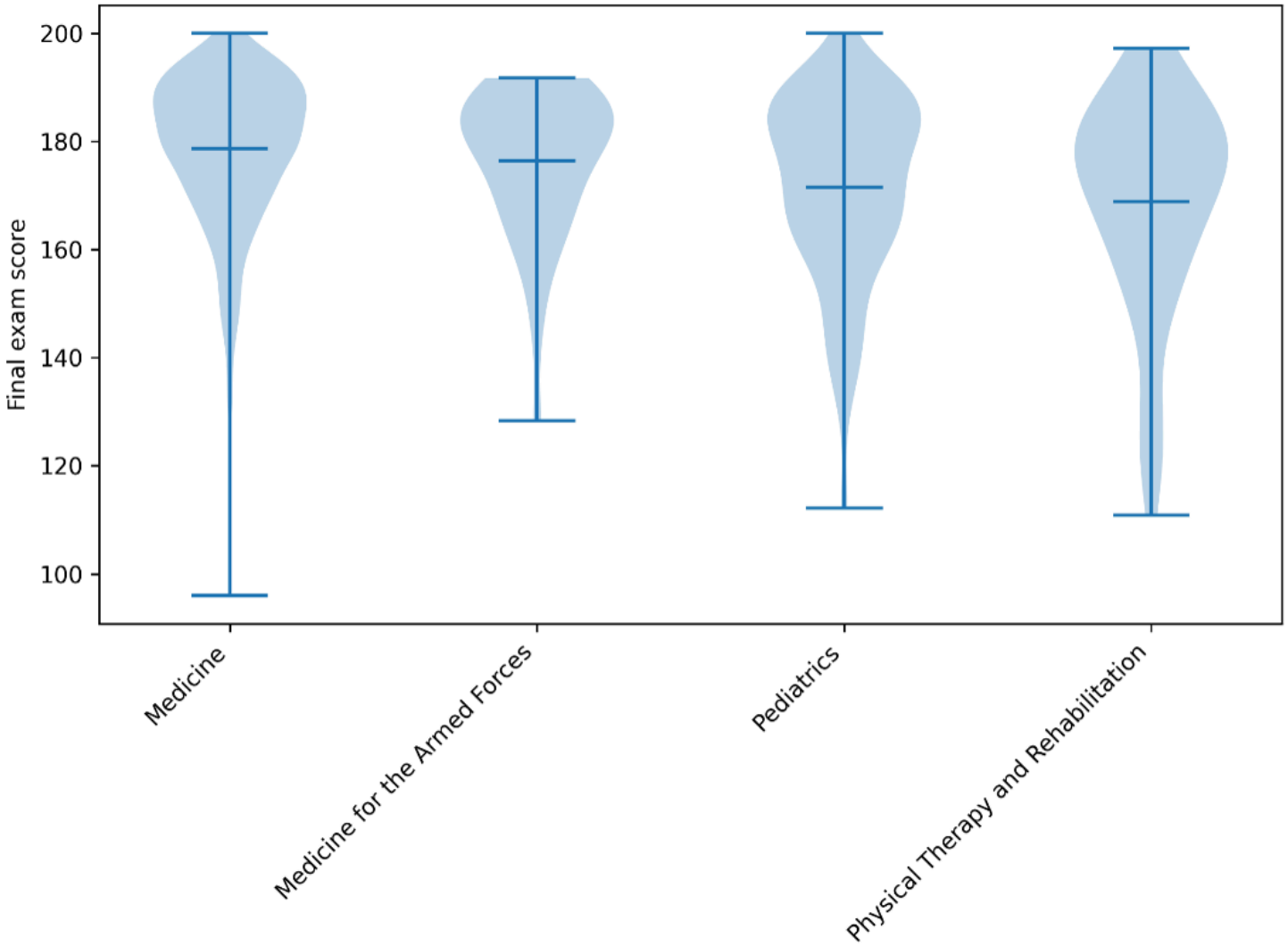


Figure 2

Overall exam performance. Medicine n=989, Medicine for the Armed Forces n=40, Pediatrics n=137, Physical Therapy and Rehabilitation n=67.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx