AI-enhanced social robotic versus computer-based virtual patients: a quantitative comparison of self-perceived clinical reasoning in undergraduate medical education

Alexander Borg, Jonathan Schiött, William Ivegren, Cidem Gentline, Viking Huss, Anna Hugelius, Benjamin Jobs, Fabricio Espinosa, Mini Ruiz, Samuel Edelbring, Carina Georg, Gabriel Skantze, Ioannis Parodis

Submitted to: Journal of Medical Internet Research
on: August 17, 2025

Table of Contents

AI-enhanced social robotic versus computer-based virtual patients: a quantitative comparison of self-perceived clinical reasoning in undergraduate medical education

Alexander Borg[1*]; Jonathan Schiött[1]; William Ivegren[1]; Cidem Gentline[1*]; Viking Huss[2]; Anna Hugelius[1]; Benjamin Jobs[1]; Fabricio Espinosa[1]; Mini Ruiz[3]; Samuel Edelbring[4, 5, 6]; Carina Georg[7]; Gabriel Skantze[8]; Ioannis Parodis[1, 9]

[1] Division of Rheumatology, Department of Medicine Solna, Karolinska Institutet, Karolinska University Hospital, and Center for Molecular Medicine (CMM), Stockholm, Sweden. Stockholm SE
[2] Division of Clinical Epidemiology, Department of Medicine Solna, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden. Stockholm SE
[3] Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Stockholm, Sweden. Stockholm SE
[4] 4School of Education, Culture and Communication, Mälardalen University, Västerås, Sweden Västerås SE
[5] Unit of Teaching and Learning, Karolinska Institutet, Stockholm, Sweden. Stockholm SE
[6] School of Health Sciences, Örebro University, Örebro, Sweden. Örebro SE
[7] Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden. Stockholm SE
[8] Division of Speech Music and Hearing, Royal Institute of Technology (KTH), Stockholm, Sweden. Stockholm SE
[9] Department of Rheumatology, Faculty of Medicine and Health, Örebro University, Örebro, Sweden. Örebro SE
[*]these authors contributed equally

Corresponding Author:
Ioannis Parodis

Division of Rheumatology, Department of Medicine Solna, Karolinska Institutet, Karolinska University Hospital, and Center for Molecular Medicine (CMM), Stockholm, Sweden.
D2:01 Rheumatology, Karolinska University Hospital Solna
Stockholm
SE

Abstract

Background: Virtual Patients (VPs) are digital simulations that can be used for practicing clinical reasoning (CR) in controlled educational environments. Traditional computer-based VP platforms often lack the authenticity and interactivity that is needed for effective CR skill training. Social robotic VPs combined with artificial intelligence (AI) have shown promise in enhancing realism and engagement. However, quantitative evidence comparing this emerging technology with conventional VP platforms within the context of CR training remains limited.

Objective: The aim of this investigation was to quantitatively compare VP design elements that support CR training for medical students between an AI-enhanced social robotic VP platform and a conventional computer-based VP software, as self-reported by the students.

Methods: We conducted a randomised cross-over trial involving 178 sixth-semester medical students at Karolinska Institutet, Stockholm, Sweden. Students experienced a social robotic VP platform with an integrated large language model to support dialogue (SARI) and a conventional computer-based VP platform (VIC) during CR training sessions. VP design was evaluated using a validated questionnaire. Student preferences for CR training were assessed using categorical responses and a visual analogue scale.

Results: SARI demonstrated significant superiority over VIC across multiple domains of VP design. Students rated SARI higher for authenticity of patient encounters (median [IQR]: 4.0 [3.5–4.5] versus 3.0 [2.5–3.5]; p<0.001), professional approach during consultation (median [IQR]: 4.5 [4.0–4.8] versus 4.0 [3.5–4.5]; p<0.001), coaching during consultation (median [IQR]: 4.3 [4.0–4.7] versus 4.0 [3.7–4.7]; p<0.001), and learning effect of consultation (median [IQR]: 4.4 [4.0–5.0] versus 4.0 [3.5–4.5]; p<0.001), all on scales from 1 to 5. Students strongly favoured SARI over VIC for CR training (72% versus 14%; OR: 27.1; 95% CI: 14.3–53.7; p<0.001). These preferences were consistent in subgroup analyses of student groups of interest (female versus male students, students with or without prior VP experience, and students first introduced to SARI or VIC). Comparisons were made between SARI and VIC, SARI and equal preference, and SARI and not SARI (i.e., VIC and equal preference combined). The only comparisons not reaching statistical significance were among students with prior VP experience and those first introduced to VIC in the SARI versus not SARI comparison, although a numerical preference for SARI was still observed in these groups.

Conclusions: AI-enhanced social robotic VPs appear to offer superior design characteristics and, in our investigation, were strongly favoured by medical students for CR training compared with conventional computer-based software. These results corroborate prior research supporting the integration of AI with social robotics as a promising approach for VP simulations aimed at preparing medical students for real clinical encounters.

(JMIR Preprints 17/08/2025:82541)
DOI: https://doi.org/10.2196/preprints.82541

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

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original Paper

AI-enhanced social robotic versus computer-based virtual patients: a quantitative comparison of self-perceived clinical reasoning in undergraduate medical education

Abstract

Background: Virtual Patients (VPs) are digital simulations that can be used for practicing clinical reasoning (CR) in controlled educational environments. Traditional computer-based VP platforms often lack the authenticity and interactivity that is needed for effective CR skill training. Social robotic VPs combined with artificial intelligence (AI) have shown promise in enhancing realism and engagement. However, quantitative evidence comparing this emerging technology with conventional VP platforms within the context of CR training remains limited.
Objective: The aim of this investigation was to quantitatively compare VP design elements that support CR training for medical students between an AI-enhanced social robotic VP platform and a conventional computer-based VP software, as self-reported by the students.
Methods: We conducted a randomised cross-over trial involving 178 sixth-semester medical students at Karolinska Institutet, Stockholm, Sweden. Students experienced a social robotic VP platform with an integrated large language model to support dialogue (SARI) and a conventional computer-based VP platform (VIC) during CR training sessions. VP design was evaluated using a validated questionnaire. Student preferences for CR training were assessed using categorical responses and a visual analogue scale.
Results: SARI demonstrated significant superiority over VIC across multiple domains of VP design. Students rated SARI higher for authenticity of patient encounters (median [IQR]: 4.0 [3.5–4.5] versus 3.0 [2.5–3.5]; p<0.001), professional approach during consultation (median [IQR]: 4.5 [4.0–4.8] versus 4.0 [3.5–4.5]; p<0.001), coaching during consultation (median [IQR]: 4.3 [4.0–4.7] versus 4.0 [3.7–4.7]; p<0.001), and learning effect of consultation (median [IQR]: 4.4 [4.0–5.0] versus 4.0 [3.5–4.5]; p<0.001), all on scales from 1 to 5. Students strongly favoured SARI over VIC for CR training (72% versus 14%; OR: 27.1; 95% CI: 14.3–53.7; p<0.001). These preferences were consistent in subgroup analyses of student groups of interest (female versus male students, students with or without prior VP experience, and students first introduced to SARI or VIC). Comparisons were made between SARI and VIC, SARI and equal preference, and SARI and not SARI (i.e., VIC and equal preference combined). The only comparisons not reaching statistical significance were among students with prior VP experience and those first introduced to VIC in the SARI versus not SARI comparison, although a numerical preference for SARI was still observed in these groups.
Conclusions: AI-enhanced social robotic VPs appear to offer superior design characteristics and, in our investigation, were strongly favoured by medical students for CR training compared with conventional computer-based software. These results corroborate prior research supporting the integration of AI with social robotics as a promising approach for VP simulations aimed at preparing medical students for real clinical encounters.

Keywords: virtual patients, medical education, clinical reasoning, large language models, social robotics,

Introduction

Virtual patients (VP) are digital educational modalities that enable learners within health professions to interact with patient cases for learning purposes [1]. These educational tools complement real-life patient interactions by allowing healthcare learners to practise gathering medical histories and making clinical and diagnostic decisions in controlled, safe environments [2, 3]. VPs can be developed using various technologies and are often utilised for clinical reasoning (CR) training purposes [4, 5]. While VPs offer standardised training opportunities, medical students report that conventional computer-based implementations often lack the authenticity needed for realistic CR training, potentially limiting their educational impact [6, 7].

CR represents a fundamental cognitive process that guides diagnostic and management decisions in clinical practice [8-10]. While CR development is important for patient safety and clinical outcomes [11], and has been recommended to be explicitly addressed in health professions education curricula [12], traditional training approaches may not adequately capture the dynamic and complex nature of clinical decision-making [13]. VPs offer unique opportunities to train and assess CR skills using standardised, repeatable scenarios that simulate real clinical encounters [14, 15].

Recent advances in artificial intelligence (AI), particularly the introduction and increasing use of large language models (LLMs), have enhanced VP platforms and their ability to provide more sophisticated clinical scenarios and realistic patient responses that better simulate the complexity of real CR challenges [16-19]. When integrated with physical embodied agents such as social robotic interfaces, these platforms can deliver multimodal interactions that may enhance the cognitive engagement required for effective CR skill development [18]. Our previous qualitative research demonstrated that sixth semester medical students perceived AI-enhanced social robotic interfaces as more clinically authentic than traditional computer-based VPs for CR training [19]. While this qualitative evidence suggests that AI-enhanced social robotic VPs provide more authentic learning experiences compared to traditional computer-based platforms, no quantitative studies have hitherto examined whether these perceived advantages translate into measurably better VP design characteristics that support CR skill development.

Given the critical importance of CR skills in clinical practice and the challenges of providing authenticity using traditional VPs, quantitative evaluation is essential to determine the effectiveness of emerging VP technologies [20, 21]. This study aimed to compare VP design characteristics that facilitate medical students' training of CR skills, between an AI-enhanced social robotic and a conventional computer-based VP platform, as reported by the students.

Methods

We conducted a randomised trial with a cross-over design to compare VP design elements that support CR skill training in medical students between an LLM-empowered social robotic platform, i.e., an in-house developed Social AI-enhanced Robotic Interface (SARI) [17-19] and a computer-based VP platform, i.e., the Virtual Interactive Case system (VIC) [22]. Each student experienced both VP platforms in a randomised cross-over design, allowing them to serve as their own control and minimising between-participant variability. We collected quantitative data using a questionnaire previously developed for VP platform evaluation (measuring authenticity, professional approach, coaching quality, learning effects, and overall judgement) within the context of CR [23], combined with additional evaluation items developed by our research team to assess the preference of each VP platform for CR training. The complete questionnaire is provided in Supplementary Figure S1.

Study participants

Study participants comprised sixth-semester medical students enrolled in clinical courses at Karolinska Institutet (KI) in Stockholm, Sweden. All medical students at KI participate in clinical rotations within rheumatology at the Karolinska University Hospital as a part of their curriculum. During these clinical rotations, students participate in an educational activity called "the virtual outpatient clinic", where students meet and interact with VP cases.

Students were recruited from those undertaking these rotations between the spring term of 2024 and the spring term of 2025. All students during this period (N=421) were invited to participate in the study by completing a survey, and a total of 178 agreed to participate. There were no specific inclusion or exclusion criteria. Participation was voluntary, and students provided written informed consent prior to enrolment with the opportunity to withdraw at any time. Students received no reimbursement for participation. The Swedish Ethical Review Authority approved the study prior to initiation (registration number: 2022-04437-01).

VP case development and practice

The VP cases had been developed and implemented for the virtual outpatient clinic before the study commenced, following specific recommendations for VP case development. This included (i) structured case design models with difficulty levels adapted to students' curriculum, (ii) incorporation of CR strategies such as differential diagnostic formulation, and (iii) interactive elements in history-taking and physical examination options [24, 25]. All cases were developed in

English to also ensure accessibility for international exchange students. A total of ten cases were developed: five were presented using SARI and five using VIC. One case was identical between the two platforms, which served the comparison between the two VP platforms during the early stages of their development [17, 18]. Students interacted with five cases on one platform and four on the other, experiencing a total of nine unique VPs. The cases included representative patient scenarios of rheumatological conditions supporting the students' learning curriculum, including polymyalgia rheumatica, rheumatoid arthritis, psoriatic arthritis, ankylosing spondylitis, systemic lupus erythematosus, and Sjögren's disease. A detailed description of the case development process and the case contents has been described elsewhere [19].

The students attended the virtual outpatient clinic over a period of one and a half days, during which they encountered and interacted with all VP cases. Before training with the cases, the students received written information on generic medical history questions typically used at the rheumatology outpatient clinic to gather a structured history from patients and support diagnostic procedures (Supplementary Figure S2). The students were instructed to perform the VP sessions in pairs or small groups of three students to allow for interaction and active collaboration, which has been shown to favour CR skill training [26]. Cases were initiated with a brief case presentation, which included the patient's primary concern, age, and name. Next, the students explored the case environment in their preferred order. Cases were concluded when students perceived that they had gathered sufficient information to perform preliminary diagnostics and propose a suitable management plan for the VP. Following completion of each case, the students participated in case-specific follow-up seminars to discuss the case information and pose questions to a seminar leader, who was a consultant rheumatologist at the clinic. During these seminars, students were asked to summarise the VP case briefly in a structured manner to practise clinical communication, and to propose their recommendations for further management. These seminar series comprised two to three student groups who had just performed the same VP case (i.e., four to nine students). While all students performed all nine VP cases, half began with SARI and half began with VIC upon randomisation on the first day of the virtual outpatient clinic. Students completed the virtual outpatient clinic after finishing all VP cases along with their corresponding follow-up seminars.

Social AI-enhanced Robotic Interface (SARI)

The embodiment of SARI consisted of a social robot from Furhat Robotics, which projects an animated face onto a semi-transparent face mask. The face mask is placed on a plastic head, which is connected to a mechanical neck that allows natural head movements and adjustments of gaze direction to facilitate interaction with multiple users simultaneously, using sensors [27]. Furthermore, the robot displays facial expressions and provides affective responses along with sophisticated gaze behaviour to indicate emotions during interaction [28]. SARI combines the Furhat software development kit (FurhatSDK) with the Open AI gpt-3.5 turbo large language model (LLM) [29]. We used prompts to generate dialogue responses using specific instructions, including a detailed patient description along with the ten latest dialogue turns. The LLM was also prompted to generate facial expressions to match the emotional state of the VP during dialogue, selecting from available expressions in the FurhatSDK at specific anchor points during the dialogues [30]. A VP prompt example is provided in the Supplementary Figure S3.

Before starting a VP case in SARI, students received brief written contextual information along with results from relevant laboratory tests with their corresponding reference values.

Virtual interactive case simulator

VIC is an online computer-based VP platform where users freely interact with a VP case, while the initial presentation of the case and the conclusion remain fixed [31]. We incorporated interactive case elements primarily regarding medical history, physical examination options, laboratory tests, and relevant diagnostic imaging. There was no numerical limit to the number of questions or examination options that could be chosen. However, students were encouraged to be selective in investigating information they deemed relevant to the case as more information became available. The cases concluded with students providing preliminary diagnoses and proposed management plans by selecting from multiple-choice responses.

Data collection

Immediately after completion of the virtual outpatient clinic, students who had agreed to participate in this study completed a questionnaire to evaluate the VP platform design, with particular emphasis put on CR training. The questionnaire combined a previously developed and validated questionnaire by Huwendiek et al. [23] for VP design evaluation, using Likert-scale data based on statements regarding the VP experience, and two questions developed by our research group to ask about the preferred VP platform for CR training, for the purpose of comparisons. The complete questionnaire can be found in Supplementary Figure S1

Likert-scale data were divided into the following quantitative themes from the questionnaire by Huwendiek et al. [23]: (I) authenticity of the patient encounter and the consultation, (II) professional approach in the consultation, (III) coaching during consultation, (IV) learning effect of consultation, and (V) overall judgement of the case work-up. Within each theme, students evaluated a statement relating to their VP experience, e.g. “while working through this case, I felt I had to make the same decisions a clinician would make in real life”, which was scored from 1 (strongly disagree) to 5 (strongly agree).

The two questions relating to the preferred VP platform were: (i) “Overall, which of the platforms is preferable to you in relation to acquirement of clinical reasoning skills?”, and (ii) “On a scale from 0 to 10, where 0 is total preference for the social robot and 10 is total preference for the computer-based platform, how would you grade your preference of the virtual patient platforms compared to each other for acquirement of clinical reasoning skills?”. The first question was answered using one of three responses: SARI, VIC, or equally preferred, while the second question had the same structure as a visual analogue scale (VAS) format between the numerical values of 0 and 10, where a lower score denoted a stronger preference for SARI, a higher score denoted a stronger preference for VIC, and a score of 5 denoted equal preference.

In addition to questions relating to the VP design and evaluations of preferred platform for CR training, students provided information regarding their age and sex, whether they had any previous experience with VPs, and which platform they were first introduced to during the virtual outpatient clinic.

Statistical analysis

The Wilcoxon signed-rank test was used to compare responses to Likert-scale data from quantitative themes in the questionnaire described by Huwendiek et al. [23] and VAS responses. For each theme, individual item scores were averaged to create a composite theme score for each student. Theme scores were next compared between platforms. VAS responses from the two platforms were compared to a hypothetical score of 5 denoting equal preference. The Fisher’s exact test with Monte Carlo simulation (10.000 iterations) was used to compare frequencies of categorical responses for VP platform preference based on students’ CR training experience. Results from Wilcoxon signed-rank tests are presented as medians and the corresponding interquartile range (IQR), test statistics (W), effect size (r), and *p* value. Results from Fisher’s exact tests are presented as frequencies and the corresponding percentage, odds ratio (OR), 95% confidence interval (CI), and *p* value. The statistical analysis was performed using R version 4.3.3 (R foundation for Statistical Computing, Vienna, Austria). Differences yielding *p* values <0.05 were considered statistically significant.

Results

Demographic and study-specific characteristics

Of the 178 students who participated in the questionnaire, 93 (52%) were women and 86 (48%) were men. Most students had no previous experience with VP platforms (150 students, 84%), while 29 (16%) reported prior experience. The mean age was 25.3 (standard deviation: 5.4) years. Regarding platform order, 101 (56%) students started with SARI and 77 (43%) started with VIC.

Table 1. Comparison of median scores using Wilcoxon signed-rank test within each question of the clinical reasoning questionnaire based on Likert-scale data.

Theme	**Variable**	**SARI median [IQR]**	**VIC median [IQR]**	***W***	***r***	***p* value**
Authenticity of patient encounter						
	Q1; n=172	4.0 [3.0–4.0]	3.0 [3.0–4.0]	5306	0.34	**<0.001**
	Q2; n=173	4.0 [3.0–4.0]	3.0 [2.0–3.0]	6936	0.58	**<0.001**
Professional approach in the consultation						
	Q1; n=172	5.0 [4.0–5.0]	4.0 [3.0–5.0]	4011	0.43	**<0.001**
	Q2; n=173	4.0 [4.0–5.0]	4.0 [4.0–5.0]	1378	0.28	**<0.001**
	Q3; n=169	4.0 [3.0–5.0]	4.0 [3.0–4.0]	2146	0.43	**<0.001**
	Q4; n=171	5.0 [4.0–5.0]	4.0 [4.0–5.0]	868	0.27	**<0.001**
Coaching during consultation						
	Q1; n=174	5.0 [4.0–5.0]	5.0 [4.0–5.0]	248	0.11	0.119
	Q2; n=158	4.0 [4.0–5.0]	4.0 [3.3–5.0]	1123	0.29	**<0.001**
	Q3; n=152	4.0 [4.0–5.0]	4.0 [3.0–5.0]	1268	0.31	**<0.001**
Learning effect of consultation						
	Q1; n=175	4.0 [4.0–5.0]	4.0 [3.0–4.0]	1491	0.35	**<0.001**
	Q2; n=173	4.0 [4.0–5.0]	4.0 [3.0–4.0]	1515	0.39	**<0.001**
Overall judgement of case work-up						
	Q1; n=175	5.0 [4.0–5.0]	4.0 [4.0–5.0]	2351	0.35	**<0.001**

Data are presented as the median score (IQR), test statistic (W), and effect size (r). The total number of study participants was 178. In case of missing values, numbers of participants with available data are indicated. Statistically significant *p* values are in bold.

IQR: interquartile range; SARI: social AI-enhanced robotic interface; VIC: virtual interactive case simulator; Q: question.

Table 2. Comparison of median composite theme scores between platforms using Wilcoxon signed-rank test. Composite scores were calculated by averaging each students' responses within themes.

Theme	SARI median [IQR]	VIC median [IQR]	*W*	*r*	*p* value
Authenticity of patient encounter; n=173	4.0 [3.5–4.5]	3.0 [2.5–3.5]	9253	0.54	**<0.001**
Professional approach in the consultation; n=173	4.5 [4.0–4.8]	4.0 [3.5–4.5]	6717	0.54	**<0.001**
Coaching during consultation; n=174	4.3 [4.0–4.7]	4.0 [3.7–4.7]	4092.5	0.32	**<0.001**
Learning effect of consultation; n=175	4.4 [4.0–5.0]	4.0 [3.5–4.5]	2589	0.42	**<0.001**
Overall judgement of case work-up; n=175	5.0 [4.0–5.0]	4.0 [4.0–5.0]	2351	0.35	**<0.001**

Data are presented as the median score (IQR), test statistic (W), and effect size (r). The total number of study participants was 178. In case of missing values, numbers of participants with available data are indicated. Statistically significant *p* values are in bold.

IQR: interquartile range; SARI: social AI-enhanced robotic interface; VIC: virtual interactive case simulator; Q: question.

VP platform design evaluation for clinical reasoning training

Students consistently rated SARI as superior to VIC across multiple domains of VP design that support CR training. Results from Likert-scale data are illustrated in Figure 1, Figure 2, Table 1, and Table 2, with detailed response frequencies provided in Supplementary Tables S4–S8.

Authenticity of patient encounters

SARI demonstrated higher authenticity ratings than VIC. Students experienced that they had to make decisions as a real-life clinician to a greater degree with SARI compared to VIC (median [IQR]: 4.0 [3.0–4.0] versus 3.0 [3.0–4.0]; *W*: 5306; r: 0.34; p<0.001) and felt with SARI more like a clinician being responsible for the care of the VP (median [IQR]: 4.0 [3.0–4.0] versus 3.0 [2.0–3.0]; *W*: 6936; r: 0.58; p<0.001). The overall theme score remained significantly greater for SARI (median [IQR]: 4.0 [3.5–4.5] versus 3.0 [2.5–3.5]; *W*: 9253; r: 0.54; p<0.001).

Professional approach during consultation

Students reported significantly greater active engagement in CR processes when using SARI compared to VIC. This included more active engagement in gathering necessary clinical information (median [IQR]: 5.0 [4.0–5.0] versus 4.0 [3.0–5.0]; *W*: 4011; r: 0.43; p<0.001), revising their clinical impression as new information became available (median [IQR]: 4.0 [4.0–5.0] versus 4.0 [4.0–5.0]; *W*: 1378; r: 0.28; p<0.001), creating structured patient summaries using medical terminology (median [IQR]: 4.0 [3.0–5.0] versus 4.0 [3.0–4.0]; *W*: 2146; r: 0.43; *p*<0.001), and actively considering findings that support or refute differential diagnoses (median [IQR]: 5.0 [4.0–5.0] versus 4.0 [4.0–5.0]; *W*: 868; r: 0.27; *p*<0.001). The overall theme score was significantly greater for SARI (median [IQR]: 4.5 [4.0–4.8] versus 4 [3.5–4.5]; *W*: 6717; r: 0.54; *p*<0.001).

Coaching during consultation

Students perceived no significant difference between platforms regarding case difficulty appropriateness for their training level (median [IQR]: 5.0 [4.0–5.0] versus 5.0 [4.0–5.0]; *W*: 248; r: 0.11; *p*=0.119). However, SARI was rated significantly better in facilitating helpful interactions that enhanced diagnostic reasoning (median [IQR]: 4.0 [4.0–5.0] versus 4.0 [3.3–5.0]; *W*: 1123; r: 0.29; *p*<0.001) and for system feedback that supported diagnostic reasoning (median [IQR]: 4.0 [4.0–5.0] versus 4.0 [3.0–5.0]; *W*: 1268; r: 0.31; *p*<0.001). The overall theme score favoured SARI (median [IQR]: 4.3 [4.0–4.7] versus 4.0 [3.7–4.7]; *W*: 4092.5; r: 0.32; *p*<0.001).

Learning effect of consultation

Students felt significantly better prepared for real clinical encounters after having used SARI compared to VIC. This included feeling better prepared to confirm the diagnosis and exclude differential diagnoses in real patients with similar complaints (median [IQR]: 4.0 [4.0–5.0] versus 4.0 [3.0–4.0]; *W*: 1491; r: 0.35; *p*<0.001) and feeling better prepared to provide care for real patients with similar conditions (median [IQR]: 4.0 [4.0–5.0] versus 4.0 [3.0–4.0]; *W*: 1515; r: 0.39; *p*<0.001). The overall theme score significantly favoured SARI (median [IQR]: 4.0 [4.0–5.0] versus 4.0 [3.5–4.5]; *W*: 2589; r: 0.42; *P*<0.001).

Overall judgement

Students rated the interaction with SARI as a significantly more worthwhile learning experience compared to VIC (median [IQR]: 5.0 [4.0–5.0] versus 4.0 [4.0–5.0]; *W*: 2351; r: 0.35; *p*<0.001). However, both platforms received generally positive evaluations from the students.

VP platform preference for CR training

Visual analogue scale ratings

Students reported an overall strong preference for SARI over VIC for CR training (median: 3.0; IQR: 2.0–5.0; *W*: 1604.5; r: 0.60; *p*<0.001). This preference pattern remained statistically significant across all subgroups of interest, i.e., if students were female or male, if they had or did not have prior experience with VPs, and if they had started with SARI or VIC (Figure 3 and Supplementary Tables S9–S12).

Categorical preference responses

When asked to choose their preferred platform, students strongly favoured SARI over VIC (72% versus 14%; OR: 27.1; 95% CI: 14.3–53.7; p<0.001). The preference for SARI over equal preference was also statistically significant (72% versus 15%; OR: 23.1; 95% CI: 8.5–54.6; p<0.001). When comparing SARI to VIC or equal preference combined, SARI remained strongly favoured (72% versus 28%; OR: 6.3; 95% CI: 3.9–10.4; P<0.001).

Similar patterns were observed across student subgroups in most comparisons. However, the preference difference did not reach statistical significance in the comparison between SARI and VIC or equal preference combined among students with prior VP experience (62% versus 38%; OR: 2.6; 95% CI: 0.8–8.9; p=0.114) and those introduced to VIC first (55% versus 45%; OR: 1.5; 95% CI: 0.7–2.9; p=0.327), although a numerical preference for SARI was still evident in both groups. Results are illustrated in Figure 4.

Discussion

Principal findings

This randomised cross-over trial provides the first quantitative evidence that AI-enhanced social robotic VPs offer significant advantages over conventional computer-based platforms for CR training in medical students. SARI demonstrated consistent superiority across multiple validated domains of VP design, with students overwhelmingly preferring the AI-enhanced social robotic interface for CR training. The effect sizes observed (r ranging from 0.27 to 0.60) represent moderate to large effects according to Cohen's conventions, with the strongest effect seen for overall platform preference (r=0.60). While the Likert-scale ratings showed moderate advantages for design characteristics in SARI, the categorical preference data revealed much stronger student preference (OR: 27.1), suggesting that the overall educational experience of SARI extends beyond individual distinct design elements captured by the Likert-scales of the study questionnaire.

Students perceived SARI as significantly more authentic than VIC in terms of feeling like they had to make decisions as real-life clinicians and feeling like the clinician being responsible for the care of the patient. Importantly, according to educational theory, authentic learning environments are crucial for effective CR development [14]. Our findings validate previous qualitative research from our group, which also demonstrated that SARI yields greater perceived authenticity compared to VIC [19]. The physical embodiment, combined with natural gaze behaviour, facial expressions, and conversational AI, appears to create a more immersive clinical simulation that better prepares students for real patient encounters.

The superiority of SARI in promoting professional approach during consultations suggests that social robotic embodiment facilitated more active cognitive engagement in CR processes. Students consistently reported being more actively engaged in gathering clinical information, revising their clinical impressions, creating structured patient summaries, and considering differential diagnoses when using SARI compared to VIC. This finding is particularly significant because effective VP platforms should specifically support active cognitive engagement rather than passive information processing [32, 33]. The multimodal nature of SARI in combining visual, auditory, and conversational elements may activate multiple cognitive pathways that enhance learning retention and skill transfer [17-19].

The advantages of SARI in providing helpful interactions and feedback that enhanced diagnostic reasoning suggest that AI-enhanced conversational interfaces can provide more adaptive and contextually appropriate educational support than traditional platforms. Students felt better prepared for real clinical encounters after having used SARI compared to VIC, indicating potential for improved skill transfer from virtual to actual clinical practice. This finding addresses a critical challenge in medical education which involves ensuring that VP experiences translate into improved real-world clinical competencies [34].

The superiority of SARI was generally robust across student subgroups. However, two notable exceptions warrant consideration. Students with prior VP experience showed only a numerical preference for SARI that did not reach statistical significance in the comparison with VIC or equal preference combined, suggesting that familiarity with VP technology may attenuate the perceived benefits of a new educational technology, such as the AI-enhanced social robotic interface introduced in this study. Additionally, students who were first introduced to VIC showed a smaller preference margin for SARI that did not reach statistical significance in the comparison between SARI and VIC or equal preference combined, possibly highlighting the importance of first impressions, which may carry more weight than subsequent exposures. This finding aligns with the earlier observation that familiarity may influence the perceived benefits of a new technology. Taken together, these findings indicate that while SARI generally offers advantages, perceived benefits may be influenced by students' prior exposure to similar technologies, as well as the order in which technologies are encountered. While the cross-over design mitigated learning effects, some degree of order effect likely remained, as complete elimination is inherently challenging.

Our results extend previous research on VPs in medical education by demonstrating quantifiable advantages of AI-enhanced social robotics. A recent systematic review identified the need for VP platforms that specifically target CR components such as problem presentation, hypothesis generation, and diagnostic justification [35]. Our findings suggest that AI-enhanced social robotic platforms may be well-suited for addressing these educational needs through their ability to provide dynamic, responsive interactions that mirror real clinical encounters more closely than traditional computer-based systems.

It is important to note that our findings were not uniformly consistent across all subgroups: students with prior VP experience and those first exposed to VIC demonstrated only numerical, non-significant preferences for SARI when compared with VIC or equal preference combined. While this may partly reflect limitations in statistical power, individual learner characteristics and contextual implementation factors may also impact students' perceptions of the effectiveness of new educational modalities. Long-term studies examining sustained educational benefits would strengthen the evidence for AI-enhanced VP platforms. We would also like to emphasise that the two VP platforms incorporate different design features, and that the purpose of this study was to explore students' perspectives on these design elements rather than evaluate the specific contributions of each platform to distinct aspects of CR, especially since no assessment of implemented CR skills was undertaken. Future research could further investigate the added value of platform-specific differences for CR training outcomes in implementation settings, for example through examination-based evaluations

Limitations and strengths

Several limitations warrant consideration when interpreting our results. Firstly, our findings are based on students' perceptions of VP design elements rather than objective measures of CR skill acquisition. Future research that incorporates validated CR assessments and longitudinal evaluation of skill transfer to real clinical encounters is warranted [36]. While students consistently favoured SARI, it is important to acknowledge that positive reception of novel technology does not always correlate with improved educational outcomes [37]. Secondly, the single-centre and discipline-specific design within rheumatology may limit the generalisability of the findings to other educational contexts or medical specialities. Thirdly, the use of English in case interaction rather than the students' native language (Swedish) may have affected perceived authenticity and interaction quality, though this limitation likely affected both platforms similarly.

Our study also has several strengths. The randomised cross-over design allowed students to serve as their own controls, minimising between-participant variability. We used a validated questionnaire specifically designed for VP design evaluation with emphasis on CR training. The study included a substantial sample (n=178) representing 42% of eligible students, which supports the generalisability of the findings within the target population. Furthermore, this is the first quantitative comparison between an AI-enhanced social robotic VP platform and a conventional computer-based VP software for CR training within medical education.

Implications

The findings of the present study have important implications for curricular development within medical education. AI-enhanced social robotic VP platforms may be considered superior alternatives to conventional computer-based systems, particularly for CR training. However, implementation decisions should also account for economic factors, technical infrastructure requirements, and faculty training needs. The integration of LLMs with social robotics represents a significant technological advancement that may justify investment despite potentially higher initial costs compared to traditional platforms. Over a longer term, the initial higher cost may limit costs from medical errors, through contribution to better preparation of students in advanced yet safe patient simulation environments.

Conclusions

This study provides robust quantitative evidence that AI-enhanced social robotic VP platforms offer significant advantages over conventional computer-based platforms for CR training within medical education. The consistent superiority of SARI across multiple domains of VP design, combined with strong student preferences regardless of individual characteristics, suggests that embodied AI platforms represent a meaningful advancement in medical pedagogy technology. These findings support the integration of LLMs with social robotics as a promising approach for developing more effective VP simulations that better prepare medical students for real clinical encounters.

Acknowledgements

The authors would like to thank the medical students who participated in the study, as well as the medical staff at the Division of Rheumatology at the Karolinska University Hospital, Stockholm, Sweden.

Authors’ contributions

Conception and design of the work: AB, MR, SE, CaG, GS, IP. Acquisition of data: AB, JS, WI, CiG, VH, AH, BJ, FE, IP. Statistical analysis and interpretation of data: AB, MR, SE, CaG, GS, IP. Original draft: AB, IP. Critical revision of the manuscript for important intellectual content: all authors. All authors reviewed and approved the final version of the manuscript prior to submission and agree to be accountable for all aspects of the work.

Conflicts of Interest

IP has received research funding and/or honoraria from Amgen, AstraZeneca, Aurinia, BMS, Elli Lilly, Gilead, GSK, Janssen, Novartis, Otsuka, and Roche. GS is a co-founder and Chief Scientist at Furhat Robotics. The other authors declare that they have no conflicts of interest.

Data availability

The datasets generated and analysed during this study are available from the corresponding author upon reasonable request. Access to data will be granted following appropriate ethical review and data sharing agreements.

Abbreviations

CI: confidence interval

CR: clinical reasoning

IQR: interquartile range

KI: Karolinska Institutet

LLM: large language model

OR: odds ratio

PMR: polymyalgia rheumatica

SARI: Social AI-enhanced Robotic Interface

SD: standard deviation

VAS: visual analogue scale

VIC: Virtual Interactive Case system

VP: virtual patient

Figure legends

Figure 1. Violin plots illustrating results from Wilcoxon signed-rank test on Likert-scale data from student evaluations of VP platform design with an emphasis on CR training, comparing SARI (blue colour) and VIC (red colour). Violin plots show score distributions; white boxes indicate the median and IQR, and numbers indicate mean scores. Panel A denotes theme A: authenticity of patient encounter (question 1: I felt I had to make the same decisions a doctor would make in real life; question 2: I felt I was the doctor caring for this patient). Panel B denotes theme B: professional approach in consultation (question 1: I was actively engaged in gathering the information I needed; question 2: I was actively engaged in revising my reasoning as new information became available; question 3: I was actively engaged in creating a summary of the patient’s problems using medical terms; question 4: I was actively engaged in thinking about which findings supported or refuted each diagnosis I was considering). Panel C denotes theme C: coaching during consultation (question 1: I felt that the case was at the appropriate level of difficulty for my training; question 2: the questions I was asked while working through this case helped me to learn; question 3: the feedback I received was helpful in enhancing my diagnostic reasoning process). Panel D denotes theme D: learning effect of consultation (question 1: I feel better prepared to confirm a diagnosis and exclude differential diagnoses in a real patient with this complaint; question 2: I feel better prepared to care for a real patient with this condition). CR: clinical reasoning; IQR: interquartile range; ns: non-significant; SARI: Social AI-enhanced Robotic Interface; VIC: Virtual Interactive Case system; VP: virtual patient; Q: question.

Figure 2. Violin plots illustrating results from Wilcoxon signed-rank test on Likert-scale data from student evaluations of VP platform design with an emphasis on CR training regarding theme overall judgement, comparing SARI (blue colour) and VIC (red colour). The item reads: Overall, working through this case was a worthwhile learning experience. Violin plots show score distributions; white boxes indicate the median and IQR, and numbers indicate mean scores. CR: clinical reasoning; IQR: interquartile range; ns: non-significant; SARI: Social AI-enhanced Robotic Interface; VIC: Virtual Interactive Case system; VP: virtual patient; Q: question.

Figure 3. Density plots illustrating distributions of responses regarding VP platform preference. Shown are also results from Wilcoxon signed-rank tests performed for comparisons of scores with a hypothetical score of 5 (equal preference of platforms) for each student. The different horizontal

panels show, from top to bottom, the overall distribution of responses in the entire cohort of students, overlayed distributions in women and men, overlayed distributions in students with and without prior experience of VPs, and overlayed distributions in subgroups of students starting with SARI or VIC. SARI: Social AI-enhanced Robotic Interface; VIC: Virtual Interactive Case system; VP: virtual patient.

Figure 4. Forest plots illustrating results from Fisher's exact test with Monte Carlo simulation (10,000 iterations) on proportion of students preferring SARI versus comparator. Panel A shows comparisons between SARI and VIC (blue colour). Panel B illustrates comparisons between SARI and Equal preference (green colour). Panel C shows comparisons between SARI and Not SARI (VIC or equal preference combined; red colour). Circles denote odds ratios (ORs) and whiskers denote 95% confidence intervals (CIs) on a logarithmic scale. SARI: Social AI-enhanced Robotic Interface; VIC: Virtual Interactive Case system.

References

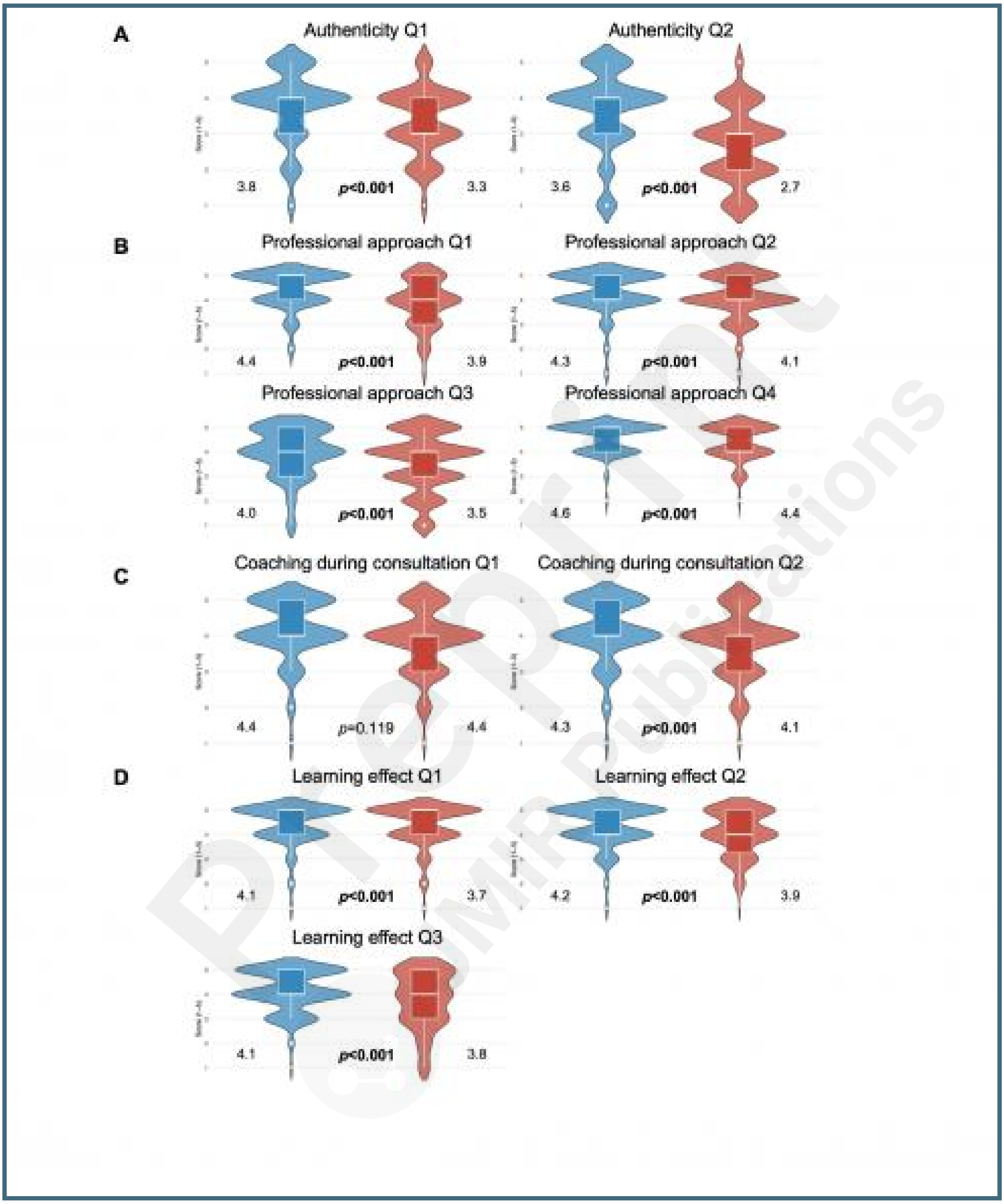
1. Ellaway R, Candler C, Greene P, Smothers V. An architectural model for MedBiquitous virtual patients. Baltimore, MD: MedBiquitous. 2006;6.
2. Consorti F, Mancuso R, Nocioni M, Piccolo A. Efficacy of virtual patients in medical education: a meta-analysis of randomized studies. Computers & Education. 2012;59(3):1001-8.
3. Yardley S, Teunissen PW, Dornan T. Experiential learning: transforming theory into practice. Medical teacher. 2012;34(2):161-4.
4. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. Med Educ. 2009 Apr;43(4):303-11. PMID: 19335571. doi: 10.1111/j.1365-2923.2008.03286.x.
5. Ellaway R. Weaving the 'e's together. Med Teach. 2006 Nov;28(7):587-90. PMID: 17594547. doi: 10.1080/01421590600909070.
6. Edelbring S, Dastmalchi M, Hult H, Lundberg IE, Dahlgren LO. Experiencing virtual patients in clinical learning: a phenomenological study. Advances in health sciences education. 2011;16:331-45.
7. Huwendiek S, Reichert F, Bosse HM, De Leng BA, Van Der Vleuten CP, Haag M, et al. Design principles for virtual patients: a focus group study among students. Medical education. 2009;43(6):580-8.
8. Young M, Thomas A, Lubarsky S, Ballard T, Gordon D, Gruppen LD, et al. Drawing boundaries: the difficulty in defining clinical reasoning. Academic Medicine. 2018;93(7):990-5.
9. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005 Jan;39(1):98-106. PMID: 15612906. doi: 10.1111/j.1365-2929.2004.01972.x.
10. Higgs J, Jensen GM, Loftus S, Trede FV, Grace S. Clinical Reasoning in the Health Professions E-Book: clinical Reasoning in the Health Professions E-Book: Elsevier Health Sciences; 2024. ISBN: 0443111650.
11. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of internal medicine. 2005;165(13):1493-9.

12. Parodis I, Andersson L, Durning SJ, Hege I, Knez J, Kononowicz AA, et al. Clinical Reasoning Needs to Be Explicitly Addressed in Health Professions Curricula: Recommendations from a European Consortium. Int J Environ Res Public Health. 2021 Oct 25;18(21). PMID: 34769721. doi: 10.3390/ijerph182111202.
13. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Academic Medicine. 2019;94(6):902-12.
14. Cook DA, Erwin PJ, Triola MM. Computerized Virtual Patients in Health Professions Education: A Systematic Review and Meta-Analysis. Academic Medicine. 2010;85(10):1589-602. PMID: 00001888-201010000-00016. doi: 10.1097/ACM.0b013e3181edfe13.
15. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. Journal of medical Internet research. 2019;21(7):e14676.
16. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Medical Teacher. 2024;46(4):446-70.
17. Borg A, Parodis I, Skantze G. Creating Virtual Patients using Robots and Large Language Models: A Preliminary Study with Medical Students. Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction; Boulder, CO, USA: Association for Computing Machinery; 2024. p. 273–7.
18. Borg A, Georg C, Jobs B, Huss V, Waldenlind K, Ruiz M, et al. Virtual Patient Simulations Using Social Robotics Combined With Large Language Models for Clinical Reasoning Training in Medical Education: Mixed Methods Study. J Med Internet Res. 2025;27:e63312. doi: 10.2196/63312.
19. Borg A, Jobs B, Huss V, Gentline C, Espinosa F, Ruiz M, et al. Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology. Rheumatology International. 2024 2024/12/01;44(12):3041-51. doi: 10.1007/s00296-024-05731-0.
20. Cohen A, Sur M, Weisse M, Moffett K, Lancaster J, Saggio R, et al. Teaching diagnostic reasoning to faculty using an assessment for learning tool: Training the trainer. MedEdPORTAL. 2020;16:10938.
21. Gruppen LD. Clinical reasoning: defining it, teaching it, assessing it, studying it. Western Journal of Emergency Medicine. 2016;18(1):4.
22. Virtual interactive case system. [cited 2024 17 apr]; Available from: https://pie.med.utoronto.ca/VIC/index.htm.
23. Huwendiek S, De Leng BA, Kononowicz AA, Kunzmann R, Muijtjens AMM, Van Der Vleuten CPM, et al. Exploring the validity and reliability of a questionnaire for evaluating virtual patient design with a special emphasis on fostering clinical reasoning. Med Teach. 2015 Aug;37(8):775-82. PMID: 25313931. doi: 10.3109/0142159X.2014.970622.
24. Posel N, Fleiszer D, Shore BM. 12 Tips: Guidelines for authoring virtual patient cases. Med Teach. 2009 Aug;31(8):701-8. PMID: 19513927. doi: 10.1080/01421590902793867.
25. Posel N, Mcgee JB, Fleiszer DM. Twelve tips to support the development of clinical reasoning skills using virtual patient cases. Medical teacher. 2015;37(9):813-8.
26. Edelbring S, Parodis I, Lundberg IE. Increasing reasoning awareness: video analysis of students' two-party virtual patient interactions. Jmir Med Educ. 2018;4(1):e9137.
27. Al Moubayed S, Beskow J, Skantze G, Granström B, editors. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011,

Revised Selected Papers; 2012: Springer.
28. Mishra C, Offrede T, Fuchs S, Mooshammer C, Skantze G. Does a robot's gaze aversion affect human gaze aversion? Frontiers in Robotics and AI. 2023;10.
29. OpenAI. ChatGPT: Optimizing Language Models for Dialogue. 2022 [cited 2024 06/04]; Available from: https://openai.com/blog/chatgpt/.
30. Irfan B, Kuoppamäki S-M, Skantze G. Between Reality and Delusion: Challenges of Applying Large Language Models to Companion Robots for Open-Domain Dialogues with Older Adults. 2023.
31. Huwendiek S, De Leng BA, Zary N, Fischer MR, Ruiz JG, Ellaway R. Towards a typology of virtual patients. Medical teacher. 2009;31(8):743-8.
32. Hege I, Kononowicz AA, Adler M. A Clinical Reasoning Tool for Virtual Patients: Design-Based Research Study. Jmir Med Educ. 2017 Nov 2;3(2):e21. PMID: 29097355. doi: 10.2196/mededu.8100.
33. Gordon D, Rencic JJ, Lang VJ, Thomas A, Young M, Durning SJ. Advancing the assessment of clinical reasoning across the health professions: Definitional and methodologic recommendations. Perspect Med Educ. 2022 Mar;11(2):108-14. PMID: 35254653. doi: 10.1007/s40037-022-00701-3.
34. Jay R, Sandars J, Patel R, Leonardi-Bee J, Ackbarally Y, Bandyopadhyay S, et al. The Use of Virtual Patients to Provide Feedback on Clinical Reasoning: A Systematic Review. Acad Med. 2025 Feb 1;100(2):229-38. PMID: 39485118. doi: 10.1097/acm.0000000000005908.
35. Plackett R, Kassianos AP, Mylan S, Kambouri M, Raine R, Sheringham J. The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. BMC medical education. 2022;22(1):365.
36. Siegelman J, Bernstein L, Goedken J, Lewin L, Schneider J, Ward M, et al. Assessment of Clinical Reasoning During a High Stakes Medical Student OSCE. Perspect Med Educ. 2024;13(1):629-34. PMID: 39678073. doi: 10.5334/pme.1513.
37. Vieriu AM, Petrea G. The Impact of Artificial Intelligence (AI) on Students' Academic Development. Education Sciences. 2025;15(3):343. PMID: doi:10.3390/educsci15030343.

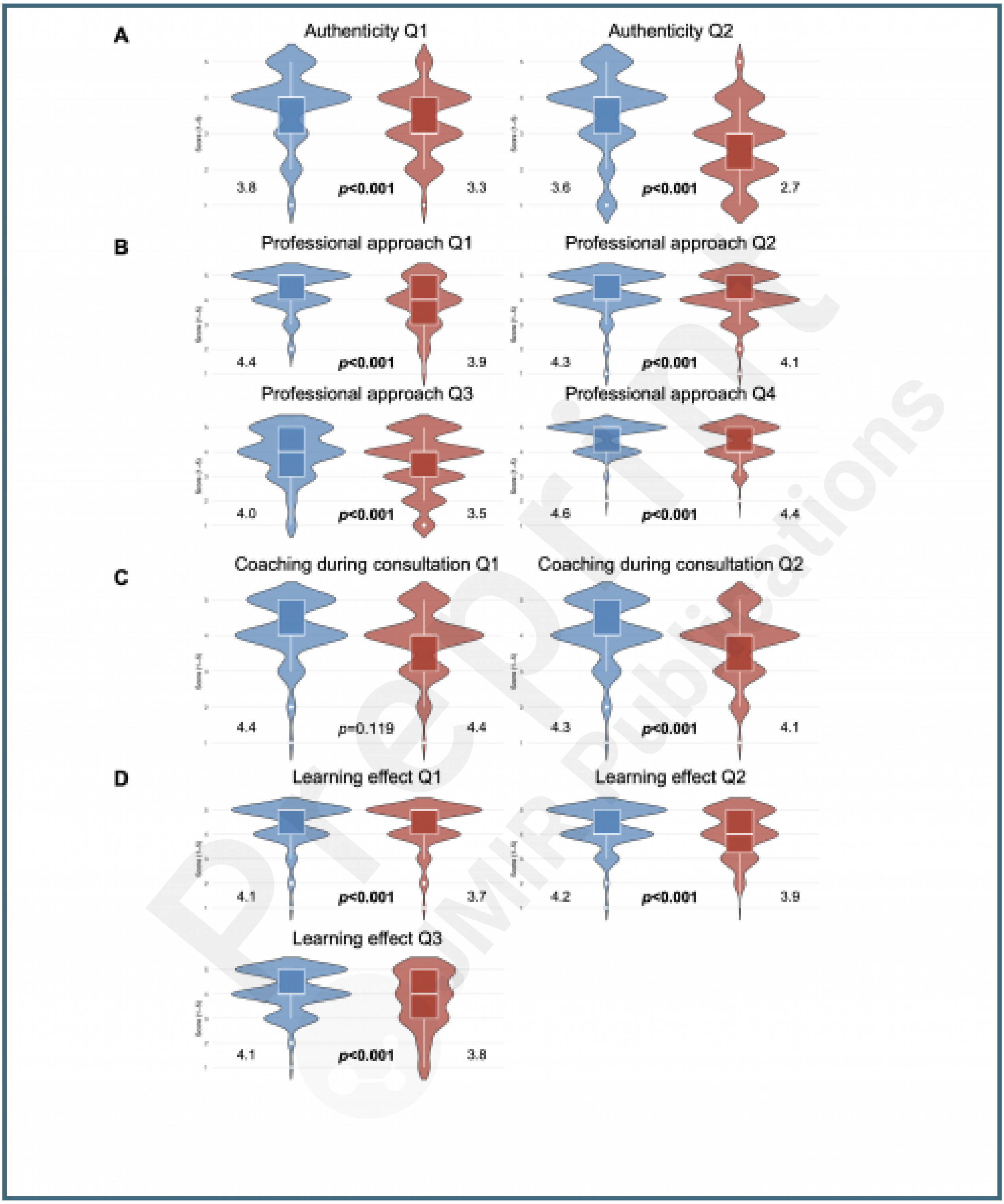
Supplementary Files

Untitled.


A
Authenticity Q1
3.8
p<0.001
3.3
Authenticity Q2
3.6
p<0.001
2.7
B
Professional approach Q1
4.4
p<0.001
3.9
Professional approach Q2
4.3
p<0.001
4.1
Professional approach Q3
4.0
p<0.001
3.5
Professional approach Q4
4.6
p<0.001
4.4
C
Coaching during consultation Q1
4.4
p=0.119
4.4
Coaching during consultation Q2
4.3
p<0.001
4.1
D
Learning effect Q1
4.1
p<0.001
3.7
Learning effect Q2
4.2
p<0.001
3.9
Learning effect Q3
4.1
p<0.001
3.8

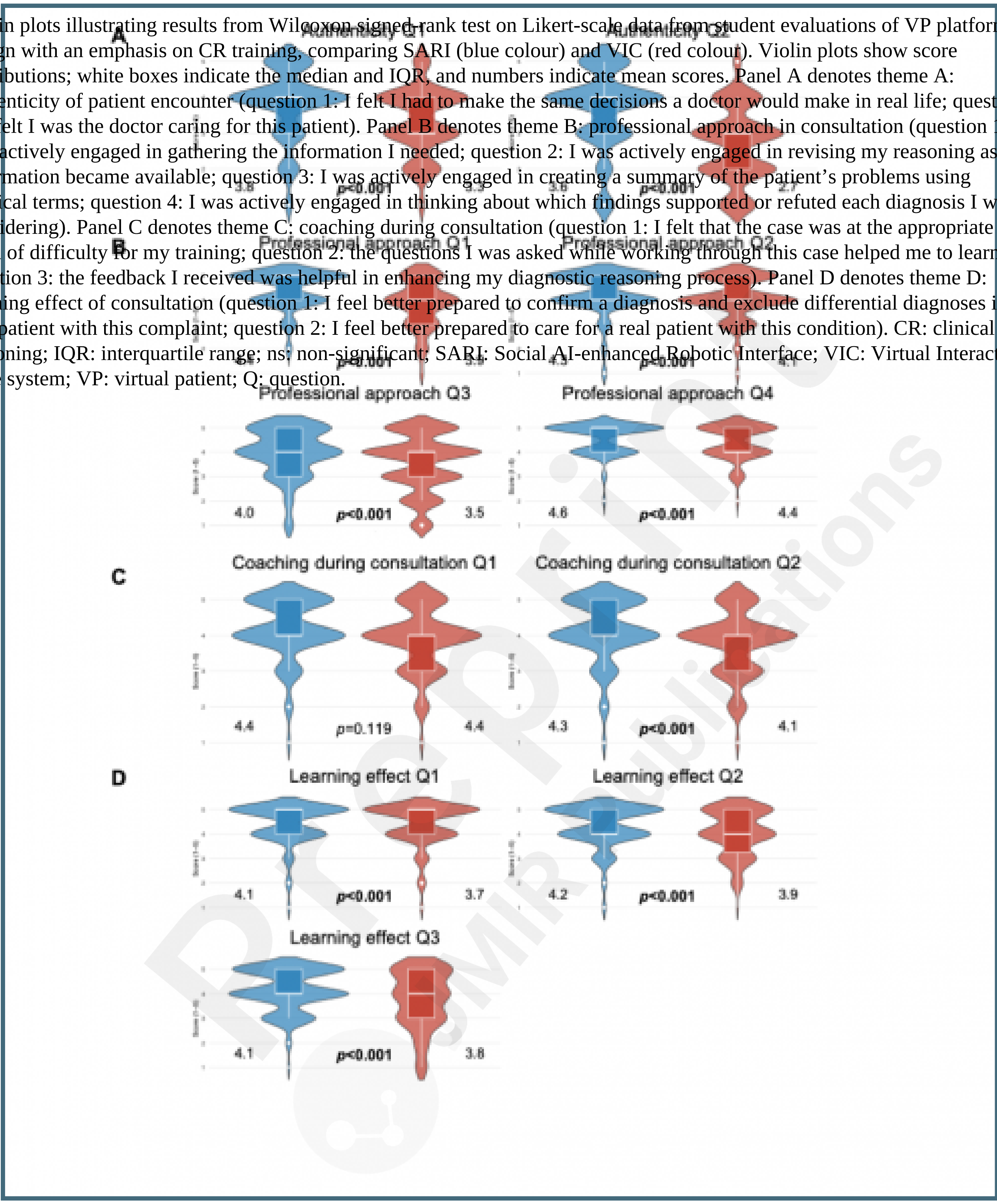
Untitled.
URL: http://asset.jmir.pub/assets/8a10e00ee63b9ca71b9ad1c6244e21c2.pdf

Untitled.

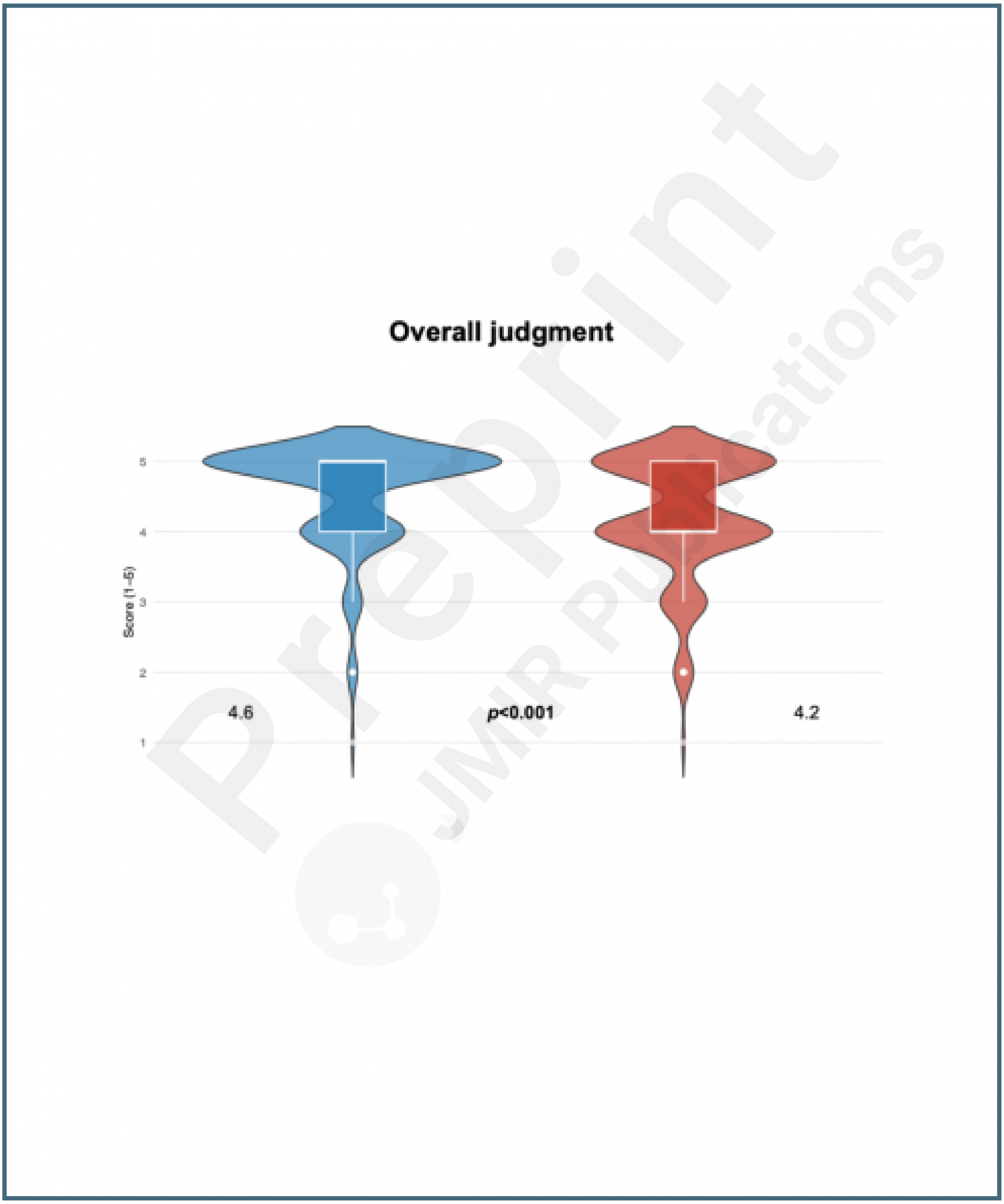

A
Authenticity Q1
3.8
p<0.001
3.3
Authenticity Q2
3.6
p<0.001
2.7
B
Professional approach Q1
4.4
p<0.001
3.9
Professional approach Q2
4.3
p<0.001
4.1
Professional approach Q3
4.0
p<0.001
3.5
Professional approach Q4
4.6
p<0.001
4.4
C
Coaching during consultation Q1
4.4
p=0.119
4.4
Coaching during consultation Q2
4.3
p<0.001
4.1
D
Learning effect Q1
4.1
p<0.001
3.7
Learning effect Q2
4.2
p<0.001
3.9
Learning effect Q3
4.1
p<0.001
3.8

Figures

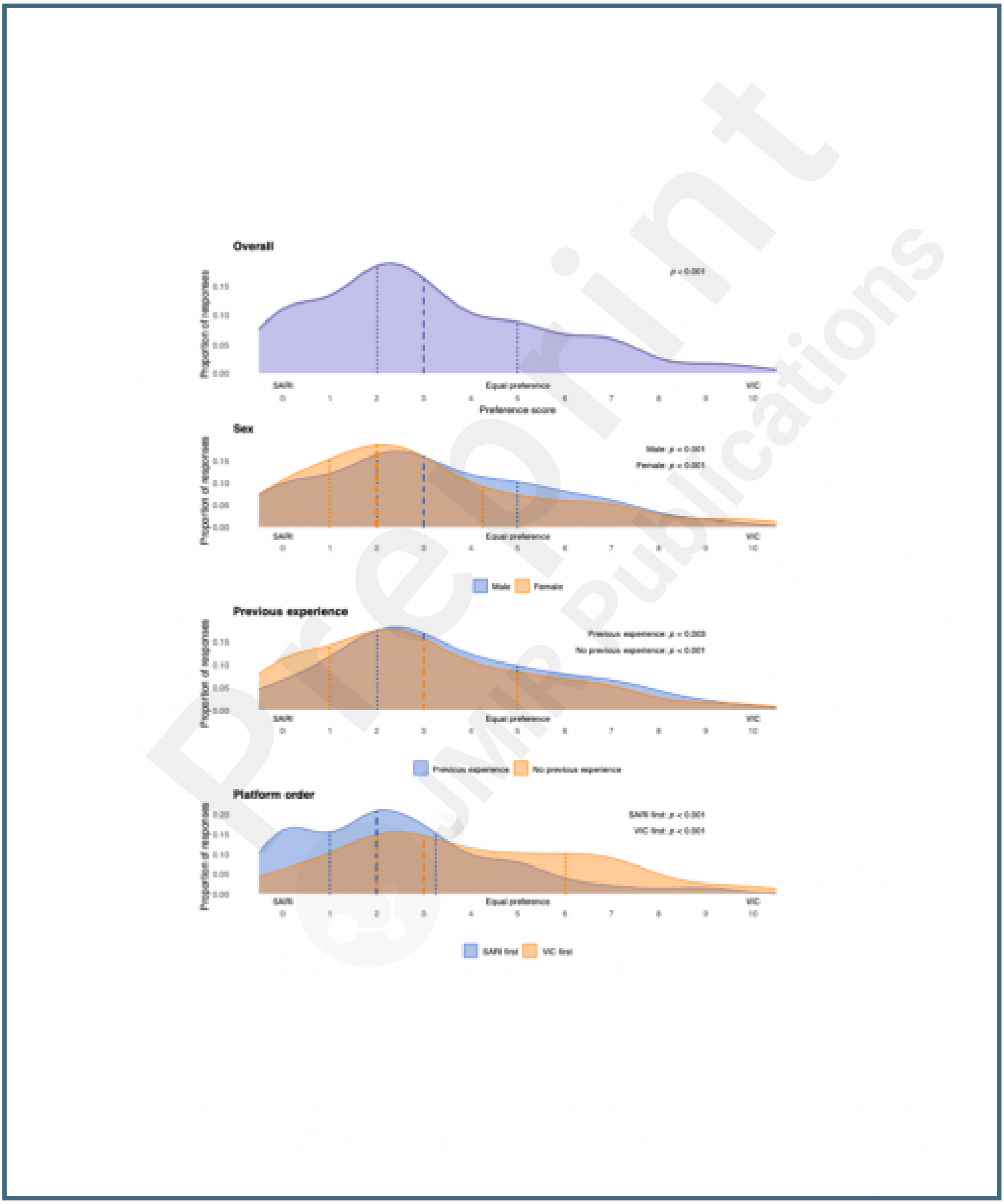
Violin plots illustrating results from Wilcoxon signed-rank test on Likert-scale data from student evaluations of VP platform design with an emphasis on CR training, comparing SARI (blue colour) and VIC (red colour). Violin plots show score distributions; white boxes indicate the median and IQR, and numbers indicate mean scores. Panel A denotes theme A: authenticity of patient encounter (question 1: I felt I had to make the same decisions a doctor would make in real life; question 2: I felt I was the doctor caring for this patient). Panel B denotes theme B: professional approach in consultation (question 1: I was actively engaged in gathering the information I needed; question 2: I was actively engaged in revising my reasoning as new information became available; question 3: I was actively engaged in creating a summary of the patient's problems using medical terms; question 4: I was actively engaged in thinking about which findings supported or refuted each diagnosis I was considering). Panel C denotes theme C: coaching during consultation (question 1: I felt that the case was at the appropriate level of difficulty for my training; question 2: the questions I was asked while working through this case helped me to learn; question 3: the feedback I received was helpful in enhancing my diagnostic reasoning process). Panel D denotes theme D: learning effect of consultation (question 1: I feel better prepared to confirm a diagnosis and exclude differential diagnoses in a real patient with this complaint; question 2: I feel better prepared to care for a real patient with this condition). CR: clinical reasoning; IQR: interquartile range; ns: non-significant; SARI: Social AI-enhanced Robotic Interface; VIC: Virtual Interactive Case system; VP: virtual patient; Q: question.



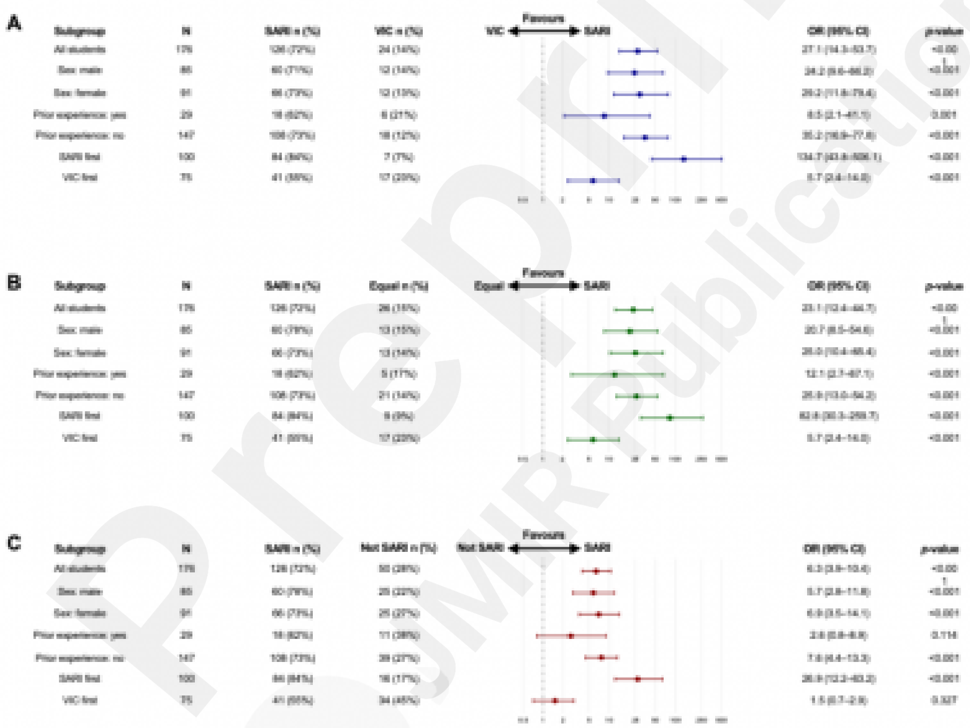
Violin plots illustrating results from Wilcoxon signed-rank test on Likert-scale data from student evaluations of VP platform design with an emphasis on CR training regarding theme overall judgement, comparing SARI (blue colour) and VIC (red colour). The item reads: Overall, working through this case was a worthwhile learning experience. Violin plots show score distributions; white boxes indicate the median and IQR, and numbers indicate mean scores. CR: clinical reasoning; IQR: interquartile range; ns: non-significant; SARI: Social AI-enhanced Robotic Interface; VIC: Virtual Interactive Case system; VP: virtual patient; Q: question.



Density plots illustrating distributions of responses regarding VP platform preference. Shown are also results from Wilcoxon signed-rank tests performed for comparisons of scores with a hypothetical score of 5 (equal preference of platforms) for each student. The different horizontal panels show, from top to bottom, the overall distribution of responses in the entire cohort of students, overlayed distributions in women and men, overlayed distributions in students with and without prior experience of VPs, and overlayed distributions in subgroups of students starting with SARI or VIC. SARI: Social AI-enhanced Robotic Interface; VIC: Virtual Interactive Case system; VP: virtual patient.



Forest plots illustrating results from Fisher’s exact test with Monte Carlo simulation (10,000 iterations) on proportion of students preferring SARI versus comparator. Panel A shows comparisons between SARI and VIC (blue colour). Panel B illustrates comparisons between SARI and Equal preference (green colour). Panel C shows comparisons between SARI and Not SARI (VIC or equal preference combined; red colour). Circles denote odds ratios (ORs) and whiskers denote 95% confidence intervals (CIs) on a logarithmic scale. SARI: Social AI-enhanced Robotic Interface; VIC: Virtual Interactive Case system.



Multimedia Appendixes

Supplementary figures, including the VP platform evaluation questionnaire, example questions for structured history taking in rheumatology, and an LLM-prompt example.
URL: http://asset.jmir.pub/assets/97fcc8188a6b85dc7070a4ffd255e292.pdf

Supplementary tables, including frequencies of responses to the VP evaluation questionnaire, and Wilcoxon signed rank test results from VAS data on student preference of VP platforms.
URL: http://asset.jmir.pub/assets/309d777e0df68ec4a413f2418ac9f986.pdf