Original Paper

The Agreement Between Virtual Patient and Unannounced Standardized Patient Assessments in Evaluating Primary Health Care Quality: Multicenter, Cross-sectional Pilot Study in 7 Provinces of China

Minrui Zeng[1], BM; Yiyuan Cai[2], PhD; Jin Cao[3], MPH; Qianyu He[1], MSN; Xiaohui Wang[3], PhD; Yun Lu[4], MM; Huijuan Liang[5], PhD; Dong Xu[6,7], PhD; Jing Liao[1], PhD

[1]Department of Medical Statistics & Epidemiology, School of Public Health, Sun Yat-sen University, Guangzhou, China

[2]Department of Epidemiology and Medical Statistics, School of Public Health, Guizhou Medical University, Guiyang, China

[3]School of Public Health, Zhejiang University School of Medicine, Hangzhou, China

[4]Department of Preventive Medicine & Maternal and Child Health, School of Public Health, Guizhou Medical University, Guiyang, China

[5]Research Institute for Health Policy of Inner Mongolia, Inner Mongolia Medical University, Hohhot, China

[6]Center for World Health Organization Studies, Department of Health Management, School of Health Management of Southern Medical University, Guangzhou, China

[7]ACACIA Lab for Implementation Research, Southern Medical University Institute for Global Health, Dermatology Hospital of Southern Medical University, Guangzhou, China

Corresponding Author:
Jing Liao, PhD
Department of Medical Statistics & Epidemiology
School of Public Health
Sun Yat-sen University
North Campus
No 74, 2nd Yat-Sen Road, Yuexiu District
Guangzhou, 510080
China
Phone: 86 13560339049
Email: liaojing5@sysu.edu.cn

Abstract

Background: The unannounced standardized patient (USP) is the gold standard for primary health care (PHC) quality assessment but has many restrictions associated with high human and resource costs. Virtual patient (VP) is a valid, low-cost software option for simulating clinical scenarios and is widely used in medical education. It is unclear whether VP can be used to assess the quality of PHC.

Objective: This study aimed to examine the agreement between VP and USP assessments of PHC quality and to identify factors influencing the VP-USP agreement.

Methods: Eleven matched VP and USP case designs were developed based on clinical guidelines and were implemented in a convenience sample of urban PHC facilities in the capital cities of the 7 study provinces. A total of 720 USP visits were conducted, during which on-duty PHC providers who met the inclusion criteria were randomly selected by the USPs. The same providers underwent a VP assessment using the same case condition at least a week later. The VP-USP agreement was measured by the concordance correlation coefficient (CCC) for continuity scores and the weighted κ for diagnoses. Multiple linear regression was used to identify factors influencing the VP-USP agreement.

Results: Only 146 VP scores were matched with the corresponding USP scores. The CCC for medical history was 0.37 (95% CI 0.24-0.49); for physical examination, 0.27 (95% CI 0.12-0.42); for laboratory and imaging tests, –0.03 (95% CI –0.20 to 0.14); and for treatment, 0.22 (95% CI 0.07-0.37). The weighted κ for diagnosis was 0.32 (95% CI 0.13-0.52). The multiple linear regression model indicated that the VP tests were significantly influenced by the different case conditions and the city where the test took place.

Conclusions: There was low agreement between VPs and USPs in PHC quality assessment. This may reflect the "know-do" gap. VP test results were also influenced by different case conditions, interactive design, and usability. Modifications to VPs and the reasons for the low VP-USP agreement require further study.

(J Med Internet Res 2022;24(12):e40082) doi: 10.2196/40082

KEYWORDS

virtual patient; unannounced standardized patient; primary health care; primary care; quality assessment; quality improvement; scenario; simulation; simulate; medical education; cross-sectional; digital health; eHealth

Introduction

Improving primary health care (PHC) services is one approach to increasing universal health coverage [1]. PHC provides comprehensive essential health care to the community by supporting access to health monitoring, diagnosis, and treatment [2] in an efficient and cost-effective manner [3,4]. PHC service quality is an important factor affecting population health outcomes and should be strengthened as part of health care system reforms [5,6] and in the face of the drastic challenges of the COVID-19 pandemic.

The unannounced standardized patient (USP) is regarded as the gold standard to assess the quality of PHC services [7-10]. USP is a rigorously trained actors portraying patients with certain diseases who anonymously visit PHC services; they can provide a standardized and timely evaluation of health care providers' performance that prevents the Hawthorne effect, that is, changes in practice associated with being observed [11]. However, USP is limited to clinical conditions that have no obvious signs and that do not require invasive examinations [12]; they are also difficult to deploy in low- and middle-income countries due to their heavy reliance on personnel and resources. The virtual patient (VP), an improvement on computerized clinical vignettes [13], has been proposed as a potential low-cost alternative to USP. VP is a software tool; they simulate real clinical scenarios and have been widely used in medical education [14] due to their low requirements for equipment, high interactivity, safety, and capacity for repeatable actions [15].

It is unknown whether assessments of quality based on VP agree with those based on USP. Prior studies mainly applied VP in medical education [16-18] as a tool to train students' clinical thinking, skills in medical history collection and diagnosis [19], and attitudes toward patients [20]. Only a few studies have directly compared VP and standardized patients; these studies have found that skills training was less effective with VPs than with standardized patients as the educational tool [21]. No study so far has used VP for PHC quality assessment. Although VP can examine users' medical knowledge (ie, their competency) in a similar way as vignettes, the results may not accurately reflect the actual performance of users in real clinical practice [22-24], and the Hawthorne effect cannot be avoided. There is some evidence that VP user interfaces and usability may influence VP-based assessment outcomes [14,25]. The extent to which a VP may serve as a quality assessment tool needs further research [26,27].

The current study belongs to a family of studies of PHC quality assessments in China based on the multicenter, nation-wide ACACIA (Health Care Quality Cohort in China) study [28]. This was a pilot study that specifically aimed to (1) examine the agreement of VPs and USPs in assessing the quality of PHC services and (2) identify factors influencing VP-USP agreement.

Methods

Study Design and Procedure

This multicenter, cross-sectional pilot study is part of the ACACIA family of studies. The ACACIA protocol has been published previously [28,29]. Briefly, ACACIA aims to develop and validate USPs and paired VPs to assess clinical quality, cost, and patient experiences in PHC across China. The study sample's representativeness was ensured by its multistage, clustered sample design [30], stratified by the average life expectancy in each province, geographic variations, and feasibility [31]. Altogether, 7 provinces were selected, and their capital cities and prefecture-level municipalities were used as a stratum; 5 townships or urban subdistricts were selected in each city based on probability proportional to size sampling. PHC facilities were then examined in each location. For this study of the agreement of VPs and USPs, a convenience sample was selected of urban PHC facilities in the capital cities of the 7 study provinces, with USP visits to these centers conducted between 2019 and 2021. All PHC providers in these centers, including licensed practicing clinicians and unlicensed clinicians under supervision of licensed physicians, served as our study population.

Altogether, 720 USP visits were conducted. On-duty PHC providers who met our criteria were randomly selected for USP visits. The PHC providers who were visited by the USPs received a VP assessment of the same cases at least a week later to prevent the practice effect [32]. The agreement between these 2 tests was analyzed with the concordance correlation coefficient (CCC) and the weighted κ.

Ethical Approval

Ethical approval was obtained from the Ethics Committee of Sun Yat-sen University (2017-007), and all PHC providers participating in the VP tests provided informed consent.

USP and VP Case Selection and Design

The USPs and VPs shared identical case designs to ensure consistency and simplify the development process. The selection and development process for these case designs was reported previously [33]. Case designs were selected based on whether the disease in question (1) had a high frequency of PHC clinical encounters, (2) had a significant disease burden, (3) was present in the main areas of PHC in China, and (4) was feasible for use in a USP test (ie, it was without obvious physiological signs and had a low risk of needing invasive tests). Twelve case designs were selected and rigorously developed: angina, asthma, diarrhea, cold, gastritis, hypertension, lower back pain, migraine, postpartum depression, stress urinary incontinence, tuberculosis, and type 2 diabetes. The validity of the case designs was verified, and they were found to have scale-level content validity indices over 0.90, role-playing fidelity over 90%, and checklist completion accuracy of 88% [34]. Most case designs had 5 modules: medical history, a physical examination, laboratory and imaging tests, diagnosis, and treatment. There were exceptions in 4 case designs (hypertension, lower back pain, migraine, and postpartum depression) that did not require laboratory or imaging tests. Due to the COVID-19 pandemic, tuberculosis cases were excluded to protect the USPs from unnecessary physical examinations, potential harm, and conflict [34]. Thus, only 11 case designs were used in this pilot study. Details of the case design development, modification, and validity testing are provided in Multimedia Appendix 1.

USP Training and Implementation

The USP actors all received at least one week of competency-based online-offline training and were assessed by specialists who were not members of the research team [34]. Before the site visits, the USP actors were further examined to ensure they could accurately portray the case designs according to the standardized training manual [30]. On the day of the visit, each USP was companied by a facilitator, who pretended to be a relative of the USP. The visits were audio recorded with a hidden recording device; these recordings were also used to monitor the performance of the USP actors and ensure checklist quality. If audio recordings were not available, field reports on what the providers said and did during the visits were upload to the online database of REDCap (Vanderbilt University) immediately after the visit to reduce recall bias. An example of a REDCap entry is provided in Multimedia Appendix 2.

VP Platform and Implementation

The VP was hosted on an online platform that could be accessed via a mobile phone or computer. The 5 VP modules used 3 different interface designs. For the medical history and diagnosis modules, the PHC providers were required to search for keywords with at least 2 characters to trigger relevant inquiries for selection. The physical examination module displayed all possible options. In the laboratory and imaging test module and the treatment module, some general options (eg, ordering blood tests or electrocardiograms) could be chosen directly, while specific options were made available after searching for keywords. All actions were recorded and uploaded online automatically.

For the field testing, PHC providers who agreed to participate in the VP tests received the VP install package for their mobile phone or personal computer alongside a user demonstration video. For each PHC provider, the cases for the VP test were the same as those for the USP test. The VPs were masked to avoid bias due to providers noticing the tested cases. The VP tests included a training VP case, which allowed the PHC providers to become familiarized with the operation of the system. There were no time limits for any of the VP tests to avoid underestimated results caused by a lack of proficiency. To facilitate the use of the VPs, some tests were organized on-site, which may have led to test results that differed from those completed by the PHC providers independently. Thus, the location and manner of the tests, as well as the number of VP tests assigned to each PHC provider and the age and sex of the providers, were recorded for analysis.

Outcome Measures

The F1 score, recall, and precision were used to measure the continuity of physical examinations, laboratory and imaging tests, and treatment [35]. However, precision and F1 score could not be calculated for medical history due to missing records for unnecessary consultations during USP visits. We used a method adapted from previous studies [36,37] in which recall represents the proportion of PHC providers who completed the checklist based on clinical guidelines, while precision was used to quantitatively assess unnecessary actions in clinical practice. The F1 score considered both recall and precision to be of equal importance and combined them. As shown in the following equation, the F1 score reflected both recall and precision:

$$F1\ score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

$$= \frac{2}{\frac{necessary\ operations}{performed\ necessary\ operations} + \frac{performed\ operations}{performed\ necessary\ operations}}$$

In the equation and in Table 1, recall represents the proportion of necessary actions that were performed in the tests and precision represents the proportion of performed actions that were necessary.

The results of the diagnoses were classified as ordinal variables in line with clinical guidelines and were rated as completely correct, partly correct, or incorrect.

Table 1. Explanation of the relationship between test results and case design for virtual patients and unannounced standardized patients. Recall is the number of performed necessary actions divided by the number of necessary actions, while precision is the number of performed necessary actions divided by the number of performed actions.

	Performed actions	Unperformed actions
Necessary actions	Number of performed necessary actions	Number of missing necessary actions
Unnecessary actions	Number of performed unnecessary actions	N/A[a]

[a] N/A: not applicable.

Statistical Analysis

Characteristics of the PHC providers and VP test information are shown as the mean (SD) for continuous variables and percentages for categorical variables. CCC, which reflects the criterion validity of the VP tests [28], was used to analyze the agreement between precision, recall, and F1 score for VP and USP tests. CCC values <0.90, 0.90 to 0.95, 0.95 to 0.99, and >0.99 were considered to represent poor, moderate, substantial, and almost perfect agreement, respectively [38]. The weighted κ (square weighted) was used to analyze diagnostic agreement [28]; weighted κ values <0.20, 0.20 to 0.40, 0.40 to 0.60, 0.60 to 0.80, and >0.80 were considered to represent poor, moderate, substantial, good, and almost perfect agreement, respectively [39,40].

Multiple linear regression was used to identify factors influencing VP-USP agreement. Using the VP tests as the dependent variable and USP tests as the independent variable, several multiple linear regression models were established, and the models were stepwise adjusted according to cases, characteristics of the PHC providers (ie, age, sex, and city), and test conditions (ie, test deployment and number of tests). Significant covariates in these models were controlled jointly in a fully adjusted model. Partial regression coefficients of the USP tests are reported. Statistical analysis was carried out using the R (version 4.0.5; R Foundation for Statistical Computing) packages stats (version 4.0.5), Desc Tools (version 0.99.43), and psych (version 2.1.9).

Results

Characteristics of PHC Providers and VP Test Information

The recruitment process is shown in Figure 1. Of 268 PHC providers who were visited by USPs, 80 agreed to conduct the VP sessions, yielding 236 valid VP scores. However, only 146 VP scores could be matched with the original USP scores.

The characteristics of the PHC providers included in the analysis were as follows: over 80% (67/80) were between 30 and 50 years old and most were male (48/80, 60%). About 40% (35/80) of the PHC providers worked in Guangzhou. Test deployment type for the VP tests was mainly field-testing (59/80, 74%) and more than half (42/80, 53%) of the PHC providers were tested by a single case. The average VP test time was 13.49 (SD 9.33) minutes. The most frequently tested case design was low back pain, with 25 tests. The least frequently tested cases were asthma, gastritis, migraine, and postpartum depression, with 7 tests each. Details are shown in Tables 2 and 3.

Figure 1. Flow chart of the recruitment process. USP: unannounced standardized patient; VP: virtual patient.

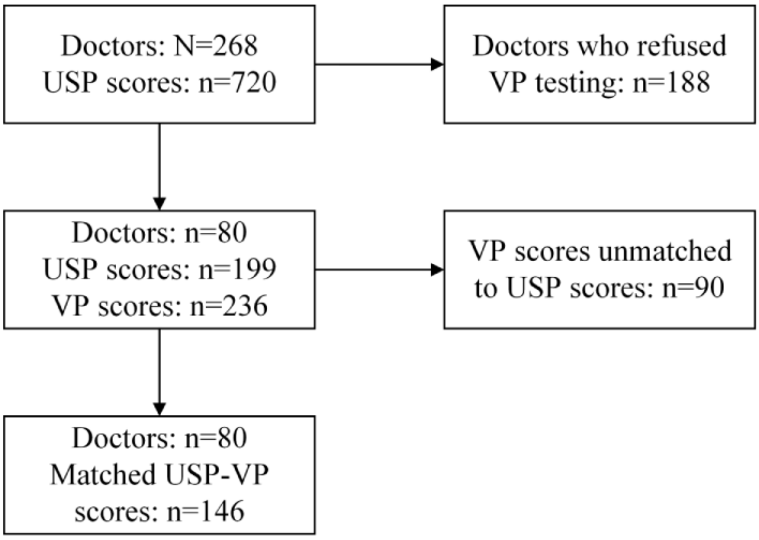

Table 2. Characteristics of primary health care providers (N=80).

Categories	Values, n (%)
Age (years)	
<30	5 (6)
30-50	67 (83)
≥50	8 (10)
Sex	
Male	48 (60)
Female	32 (40)
Location (city)	
Changsha	12 (15)
Xi'an	7 (9)
Guangzhou	35 (44)
Lanzhou	8 (10)
Hohhot	11 (14)
Guiyang	5 (6)
Chengdu	2 (3)
Case designs	
Angina	9 (6)
Asthma	7 (5)
Gastritis	7 (5)
Cold	16 (11)
Type 2 diabetes	18 (12)
Diarrhea	13 (9)
Hypertension	13 (9)
Low back pain	25 (17)
Migraine	7 (5)
Postpartum depression	7 (5)
Stress urinary incontinence	24 (16)

Table 3. Virtual patient test situations (N=80).

Categories	Values, n (%)
Test deployment	
Field-testing	59 (74)
Self-testing	21 (26)
Number of tests	
1	42 (53)
2	22 (28)
≥3	16 (20)

Agreement Between VP and USP Tests

Test outcomes and CCCs for the medical history, physical examination, laboratory and imaging tests, and treatment modules are listed in Table 4. The USP test results showed high precision (over 0.47), but the VP test results showed varying degrees of degradation (ranging from 0.25 to 0.51), which resulted in very poor agreement. Recall for medical history and treatment was similar for the USP and VP tests. It is worth noting that the physical examination module and the laboratory and imaging test module had results for the USPs that were nearly 3 times higher than for the VPs. All the CCCs for recall were poor. The F1 score and its CCC were close to the recall values, except for the CCC for physical examination. The weighted κ for diagnosis was 0.32 (95% CI 0.13-0.51), which was unsatisfactory. Details for the weighted κ are shown in Multimedia Appendix 3.

Table 4. Test outcomes and concordance correlation coefficients. Precision and F1 score could not be calculated for unannounced standardized patients for medical history due to missing consultation records.

Test modules	Precision (SD)			Recall (SD)			F1 score (SD)		
	USPs[a]	VPs[b]	CCC[c] (95% CI)	USPs	VPs	CCC (95% CI)	USPs	VPs	CCC (95% CI)
Medical history[a]	—[d]	0.51 (0.35)	—	0.19 (0.15)	0.13 (0.13)	0.37 (0.24 to 0.49)	—	0.18 (0.16)	—
Physical examination	0.47 (0.50)	0.25 (0.30)	0.13 (0.01 to 0.26)	0.11 (0.15)	0.34 (0.31)	0.04 (–0.05 to 0.13)	0.17 (0.21)	0.20 (0.19)	0.27 (0.12 to 0.42)
Laboratory and imaging tests	0.47 (0.49)	0.45 (0.34)	0.21 (0.03 to 0.38)	0.18 (0.20)	0.57 (0.36)	–0.06 (–0.15 to 0.03)	0.25 (0.27)	0.43 (0.27)	–0.03 (–0.20 to 0.14)
Treatment	0.77 (0.41)	0.45 (0.48)	0.07 (–0.06 to 0.20)	0.21 (0.18)	0.20 (0.25)	0.24 (0.10 to 0.39)	0.31 (0.23)	0.26 (0.31)	0.22 (0.07 to 0.37)

[a] USP: unannounced standardized patient.

[b] VP: virtual patient.

[c] CCC: concordance correlation coefficient.

[d] Not available.

Factors Influencing VP-USP Agreement

To explore factors that affected VP-USP agreement, we used multiple linear regression. For medical history, there was a significant correlation between VP and USP scores that remained stable after adjustment (ranging from 0.32 to 0.34, *P*<.001). In contrast, despite factor adjustment, the correlation between the VP and USP scores was not significant (*P*=.74) for the laboratory and imaging test module. The correlation was significantly weakened after adjusting the cases for the physical examination and treatment modules. Details are listed in Multimedia Appendix 4.

Using stepwise variable selection in the fully adjusted model, all the correlations between VP and USP scores became weaker after adjustment. The partial correlation coefficients were 0.314 (95% CI 0.183-0.445) for recall for the USPs for medical history; 0.071 (95% CI –0.090 to 0.023) for F1 score for physical examination; –0.025 (95% CI –0.169 to 0.118) for F1 score for laboratory and imaging tests; and 0.045 (95% CI –0.133 to 0.223) for F1 score for treatment. Furthermore, for medical history, female sex (versus male) and Changsha and Lanzhou (versus Guangzhou) were negatively associated with recall for VPs, while test time was positively associated with recall for VPs. The F1 scores for the physical examination module and the laboratory and imaging test module were only associated with case design. The F1 score for treatment was only associated with cases and the cities where the PHC providers worked. Combining the results of these models revealed that the major influencing factors were case design and city. Details are shown in Table 5.

Table 5. The association between assessments using virtual patients and unannounced standardized patients using stepwise regression for each module.

Test modules	β (95% CI)	*P* value	Standardized β
Medical history			
Recall for USPs[a]	.314 (.183 to .445)	<.001	.351
Female sex	–.049 (–.089 to –.009)	.02	–.366
City			
Guangzhou	0 (ref)		
Changsha	–.067 (–.126 to –.009)	.03	–.506
Lanzhou	–.065 (–.129 to –.001)	.049	–.489
Test time	.002 (.001 to .004)	.006	.205
Physical examination			
F1 score for USPs	.071 (–.090 to .023)	.39	.080
Case design			
Low back pain	0 (ref)		
Cold	.203 (.100 to .306)	<.001	1.086
Gastritis	.169 (.028 to .311)	.02	.907
Laboratory and imaging			
F1 score for USPs	–.025 (–.169 to .118)	.74	–.025
Case design			
Low back pain	0 (ref)		
Cold	.206 (.074 to .337)	.003	.768
Stress urinary incontinence	–.269 (–.401 to –.137)	<.001	–1.005
Type 2 diabetes	–.289 (–.416 to –.161)	<.001	–1.079
Gastritis	–.386 (–.543 to –.228)	.003	–1.440
Treatment			
F1 score for USPs	.045 (–.133 to .223)	.62	.034
Case design			
Low back pain	0 (ref)		
Cold	.432 (.300 to .563)	<.001	1.404
Hypertension	.419 (.283 to .555)	<.001	1.363
Type 2 diabetes	.350 (.223 to .477)	<.001	1.138
Postpartum depression	.173 (.003 to .344)	.05	.564
Gastritis	–.176 (–.347 to –.006)	.05	–.573
Stress urinary incontinence	–.189 (–.301 to –.078)	.001	–.615
Migraine	–.198 (–.374 to –.023)	.03	–.645
City			
Guangzhou	0 (ref)		
Lanzhou	–.190 (–.299 to –.082)	<.001	–.619
Xi'an	–.178 (–.313 to –.042)	.01	–.577

[a] USP: unannounced standardized patient.

Discussion

Principal Findings

Our study examined the agreement between using VPs and USPs to assess the quality of PHC in China. We found that the agreement between VP and USP results was low in general, which may result from the "know-do" gap. The VP tests might also have been influenced by different case conditions, different interface designs of the VPs, and the usability of the VPs.

We found that the agreement between VP and USP scores was low in our study sample. The USP scores were low in terms of recall, indicating that our study participants performed only some of the necessary actions, especially for the physical examination module and the laboratory and imaging test module. This suggests that PHC providers only partially performed the guideline-recommended checklist items in actual practice, which might be the result of a lack of incentives or limited time and resources [5,41,42]. In contrast, the VP scores showed relatively high recall scores, with module-specific variation, compared to the USP scores. One possible explanation for the low agreement is the "know-do" gap [43-45]. The USP tests assessed how the PHC providers performed in real clinical practice with evidence-based indicators [43,46]. However, the VP testing was more likely to assess whether the providers knew how to use their knowledge, which is defined as their competency [43]. With unlimited time and resources, the PHC providers might have performed extra actions to exclude other diseases, even more than required for differential diagnosis. Thus, VP testing may be more akin to examinations in a medical training setting and therefore indicate the competence of the examinees, whereas USP testing may be more likely to assess the quality of care in actual practice [17,47]. Similar findings have been reported by previous studies of medical education, which found that examinees in VP tests tended to explore all possible information [14]. Therefore, despite our intention to use the interactive VPs as an alternative to USPs as a quality assessment tool [17], our current findings on the agreement between VP and USP test results are insufficient to provide strong evidence for such a substitution.

Further analysis using multiple linear regression suggested that VP-based performance varied with case design, indicating that PHC providers' competency differed with clinical case design. Specifically, variations for type 2 diabetes and gastritis were observed for the test modules and low scores were observed for laboratory and imaging tests for both case designs, while higher scores were seen for treatment for type 2 diabetes than for gastritis. This finding may indicate that PHC providers were more familiar with type 2 diabetes, which is commonly seen in PHC and has a distinctive medical history and physical signs. As a result, laboratory and imaging tests for type 2 diabetes were more likely to be omitted, while appropriate treatment was more likely to be conducted in the VP tests. In contrast, PHC providers might prefer to conduct simple physical examinations for the symptoms of abdominal pain, but they were reluctant to conduct the complex laboratory and imaging testing and treatment that should be offered in accordance with the clinical guidelines for gastritis [48].

Furthermore, the VP interface and usability also influenced the VP testing. By and large, 2 types of interface were used in the VP testing: searching for keywords for consultation for medical history and selecting from multiple choices for physical examinations (for other modules, a mixture of both interface formats was adopted). Specifically, our results showed that the recall score for medical history for the VP testing was two-thirds of the score for USP testing, while the corresponding VP score for physical examination was more than twice that of the USP score. These findings indicate that multiple choices might provide more hints, allowing PHC providers to guess a correct action more easily [14,26]. Although searching for keywords for consultation or actions leads to less bias than the hints provided by multiple choices, this interface might decrease the usability of VPs, particularly when the interaction does not provide enough options or fuzzy searches. Many of the PHC providers in the study found this interaction was not user-friendly and that the consultation questions needed were often not retrievable. Frustrated by the poor interface, they tended to suddenly end the consultation and even drop out of the VP testing entirely. Although searching for keywords decreased usability, this interface should be used for the purpose of quality assessment, albeit with modifications. The influence of the study city may reflect differences in the attitude and capacity for digital adaptation of the PHC providers; those from developed regions with wider use of digital information systems may be more receptive to digitalized medical practice than their counterparts from less-developed regions [41,49]. VP usability may also be influenced by the digital adaptation attitude of PHC providers. Previous studies have found that examinees who are more open to digital innovation, better educated, and younger are more enthusiastic about using and completing digital device–based programs [50,51] like VP testing. Due to missing information on sociodemographic characteristics, our study could not examine the statistical significance of variations in these characteristics other than the city where the participants were located. However, we did find that the agreement between VP and USP results was lower in Changsha, Lanzhou, and Xi'an.

Study Limitations

The study had several limitations. First, as a purposive sampling approach was used, our research sample may be more likely to have included PHC providers who were receptive to technological innovations, and the extent to which our findings apply to providers who are less receptive needs verification. Although our user experience analysis showed promising results, only a few participants answered the user experience questionnaire. Second, due to substantial missing data for the sociodemographic characteristics of the PHC providers, we failed to identify any remaining influence of these factors on the agreement between VP and USP scores. Third, although we found that the VP interface was a key factor influencing the VP testing, we did not perform a direct comparison of different interfaces with the same disease module. Last, we used a summary score for each module to indicate the providers' performance, assuming that individual consultation or action items had equal importance. Nevertheless, a hierarchic order may exist among consultation and action items that is specific to the disease conditions under consideration, such that a weighted score may have been better suited for quantifying the providers' performance [52].

Study Implications for Further Studies

Our findings highlight the need for further modifications to the VP platform. To improve the design of the VPs to bring them as close as possible to real clinical conditions, strict testing time limits should be implemented to enhance the sense of time pressure. Besides this, the interactive design of the VPs should opt for keyword searching over multiple choices to minimize hints. The creation of clinical settings and the application of keyword searching can be enhanced via advanced technologies, such as virtual simulation, voice input, and fuzzy retrieval [12,53], for a better, user-centered experience [54].

Moreover, to facilitate the implementation of VPs, an add-on program to widely used social software such as WeChat would be preferable to a separate application that requires installation. A short demonstration (of less than 5 minutes) of the main action steps of the VP should be embedded in the program and shown as a mandatory preview for first-time users. If needed, initial training with VP cases should also be provided, so that users can become familiar with the platform.

To better understand the agreement between VP and USP testing, future studies would benefit from systematically collecting information on potential factors contributing to differences between VPs and USPs, using both quantitative and qualitative approaches. For instance, the preferences of PHC providers for different VP interfaces could be examined with questionnaires or interviews to assess differences in perceived authenticity, cognitive load, and motivation [19]. Moreover, potential reasons for the adherence of PHC providers to clinical guidelines and associated influencing factors need to be further explored using mixed methods based on the Theoretical Domains Framework, structured questionnaires, or focus groups [55-57].

Conclusion

The agreement between VP and USP testing for PHC quality assessment was low. This low agreement may mainly reflect the "know-do" gap, while the VP test results were also influenced by different case conditions, interface design, and usability. To improve VP usability in the resource-limited settings found in PHC, VPs should be modified to be more user-centered, paying attention to the balance between enhancing usability and avoiding hints. Factors influencing the agreement between VP and USP testing need further study.

Acknowledgments

The project received support from the China Medical Board 18-300, National Natural Science Foundation of China (71974211), the Swiss Agency for Development and Cooperation (81067392), and the National Key R&D Program of China (2021ZD0113401). The authors thank the unannounced standardized patient team for the case designs and field tests and JL for proofreading the manuscript. The funders of this study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

Due to data safety and property rights, the data sets generated and analyzed during the current study are not publicly accessible but are available upon request to the corresponding author.

Authors' Contributions

JL, YC, HL, YL, XW, and DX co-designed the study. JL, DX, and YC requested the funding. JL, YC, and DX take responsibility for project administration. All authors conducted or managed the research and investigation process. MZ, YC, and JL took part in virtual patient development. MZ, YC, and JL designed the methodology. MZ, JL, YC, and JC carried out or supported the data analysis, including the statistical analyses. YC, MZ, JL, and JC take responsibility for the integrity of the data and the accuracy of the data analysis, with each author having been responsible for specific parts of the raw data set. MZ wrote the manuscript. QH and JL reviewed the manuscript. All authors critically reviewed the manuscript and contributed to the interpretation of the results. All authors had access to the data, critically reviewed the paper before publication, and take final responsibility for the decision to submit the research for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Case development, modification, and validity.
[DOCX File , 20 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Example of REDCap.
[DOCX File , 1544 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Classification of diagnosis.
[DOCX File , 18 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Regression result of different adjustment.
[DOCX File , 24 KB-Multimedia Appendix 4]

References

1. Binagwaho A, Adhanom Ghebreyesus T. Primary healthcare is cornerstone of universal health coverage. BMJ 2019 Jun 03;365:l2391. [doi: 10.1136/bmj.l2391] [Medline: 31160322]
2. Chan M. Ten years in public health, 2007–2017: report by Dr Margaret Chan, Director-General, World Health Organization. 2007. URL: https://apps.who.int/iris/bitstream/handle/10665/255355/9789241512442-eng.pdf?sequence=1&isAllowed=y [accessed 2022-11-14]
3. So CF, Chung JW. Telehealth for diabetes self-management in primary healthcare: A systematic review and meta-analysis. J Telemed Telecare 2018 Jun;24(5):356-364. [doi: 10.1177/1357633X17700552] [Medline: 28463033]
4. Kieny M, Evans DB. Universal health coverage. East Mediterr Health J 2013 Apr;19(4):305-306 [FREE Full text] [Medline: 23882953]
5. Sylvia S, Shi Y, Xue H, Tian X, Wang H, Liu Q, et al. Survey using incognito standardized patients shows poor quality care in China's rural clinics. Health Policy Plan 2015 Apr;30(3):322-333. [doi: 10.1093/heapol/czu014] [Medline: 24653216]
6. Hu R, Liao Y, Du Z, Hao Y, Liang H, Shi L. Types of health care facilities and the quality of primary care: a study of characteristics and experiences of Chinese patients in Guangdong Province, China. BMC Health Serv Res 2016 Aug 02;16(a):335 [FREE Full text] [doi: 10.1186/s12913-016-1604-2] [Medline: 27484465]
7. Rethans J, Gorter S, Bokken L, Morrison L. Unannounced standardised patients in real practice: a systematic literature review. Med Educ 2007 Jun;41(6):537-549. [doi: 10.1111/j.1365-2929.2006.02689.x] [Medline: 17518833]
8. Daniels B, Kwan A, Satyanarayana S, Subbaraman R, Das RK, Das V, et al. Use of standardised patients to assess gender differences in quality of tuberculosis care in urban India: a two-city, cross-sectional study. Lancet Glob Health 2019 May;7(5):e633-e643 [FREE Full text] [doi: 10.1016/S2214-109X(19)30031-2] [Medline: 30928341]
9. Wu Y, Zhou H, Ma X, Shi Y, Xue H, Zhou C, et al. Using standardised patients to assess the quality of medical records: an application and evidence from rural China. BMJ Qual Saf 2020 Jun;29(6):491-498 [FREE Full text] [doi: 10.1136/bmjqs-2019-009890] [Medline: 31776199]
10. Burnett AM, Lee L, McGuinness M, Varga B, Perez Hazel Y, Ho SM. Quality of refractive error care (Q.REC) in Cambodia, Malaysia and Pakistan: protocol for a cross-sectional unannounced standardised patient study. BMJ Open 2022 Mar 14;12(3):e057594 [FREE Full text] [doi: 10.1136/bmjopen-2021-057594] [Medline: 35288395]
11. Shah R, Edgar D, Evans BJW. Measuring clinical practice. Ophthalmic Physiol Opt 2007 Mar;27(2):113-125. [doi: 10.1111/j.1475-1313.2006.00481.x] [Medline: 17324200]
12. Triola M, Feldman H, Kalet AL, Zabar S, Kachur EK, Gillespie C, et al. A randomized trial of teaching clinical skills using virtual and live standardized patients. J Gen Intern Med 2006 May;21(5):424-429 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00421.x] [Medline: 16704382]
13. Peabody JW, Luck J, Glassman P, Jain S, Hansen J, Spell M, et al. Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. Ann Intern Med 2004 Nov 16;141(10):771-780 [FREE Full text] [doi: 10.7326/0003-4819-141-10-200411160-00008] [Medline: 15545677]
14. Isaza-Restrepo A, Gómez MT, Cifuentes G, Argüello A. The virtual patient as a learning tool: a mixed quantitative qualitative study. BMC Med Educ 2018 Dec 06;18(1):297 [FREE Full text] [doi: 10.1186/s12909-018-1395-8] [Medline: 30522478]
15. Rakofsky JJ, Talbot TB, Dunlop BW. A virtual standardized patient-based assessment tool to evaluate psychiatric residents' psychopharmacology proficiency. Acad Psychiatry 2020 Dec;44(6):693-700 [FREE Full text] [doi: 10.1007/s40596-020-01286-x] [Medline: 32681418]
16. Kyaw BM, Posadzki P, Paddock S, Car J, Campbell J, Tudor Car L. Effectiveness of digital education on communication skills among medical students: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Aug 27;21(8):e12967 [FREE Full text] [doi: 10.2196/12967] [Medline: 31456579]
17. Kyaw BM, Saxena N, Posadzki P, Vseteckova J, Nikolaou CK, George PP, et al. Virtual reality for health professions education: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Jan 22;21(1):e12959 [FREE Full text] [doi: 10.2196/12959] [Medline: 30668519]
18. 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. J Med Internet Res 2019 Jul 02;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]

19. Fink MC, Reitmeier V, Stadler M, Siebeck M, Fischer F, Fischer MR. Assessment of diagnostic competences with standardized patients versus virtual patients: experimental study in the context of history taking. J Med Internet Res 2021 Mar 04;23(3):e21196 [FREE Full text] [doi: 10.2196/21196] [Medline: 33661122]
20. Kron FW, Fetters MD, Scerbo MW, White CB, Lypson ML, Padilla MA, et al. Using a computer simulation for teaching communication skills: A blinded multisite mixed methods randomized controlled trial. Patient Educ Couns 2017 Apr;100(4):748-759 [FREE Full text] [doi: 10.1016/j.pec.2016.10.024] [Medline: 27939846]
21. Deladisma AM, Cohen M, Stevens A, Wagner P, Lok B, Bernard T, Association for Surgical Education. Do medical students respond empathetically to a virtual patient? Am J Surg 2007 Jun;193(6):756-760. [doi: 10.1016/j.amjsurg.2007.01.021] [Medline: 17512291]
22. Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. JAMA 2000 Apr 05;283(13):1715-1722. [doi: 10.1001/jama.283.13.1715] [Medline: 10755498]
23. Shah R, Edgar D, Evans B. A comparison of standardised patients, record abstraction and clinical vignettes for the purpose of measuring clinical practice. Ophthalmic Physiol Opt 2010 May;30(3):209-224. [doi: 10.1111/j.1475-1313.2010.00713.x] [Medline: 20444128]
24. Dresselhaus TR, Peabody JW, Luck J, Bertenthal D. An evaluation of vignettes for predicting variation in the quality of preventive care. J Gen Intern Med 2004 Oct;19(10):1013-1018 [FREE Full text] [doi: 10.1007/s11606-004-0003-2] [Medline: 15482553]
25. Schittek Janda M, Mattheos N, Nattestad A, Wagner A, Nebel D, Färbom C, et al. Simulation of patient encounters using a virtual patient in periodontology instruction of dental students: design, usability, and learning effect in history-taking skills. Eur J Dent Educ 2004 Aug;8(3):111-119. [doi: 10.1111/j.1600-0579.2004.00339.x] [Medline: 15233775]
26. Jacklin S, Maskrey N, Chapman S. Shared decision-making with a virtual patient in medical education: mixed methods evaluation study. JMIR Med Educ 2021 Jun 10;7(2):e22745 [FREE Full text] [doi: 10.2196/22745] [Medline: 34110299]
27. Hege I, Zary N, Kononowicz AA. Criteria to assess the quality of virtual patients. Stud Health Technol Inform 2012;180:954-957. [Medline: 22874334]
28. Liao J, Chen Y, Cai Y, Zhan N, Sylvia S, Hanson K, et al. Using smartphone-based virtual patients to assess the quality of primary healthcare in rural China: protocol for a prospective multicentre study. BMJ Open 2018 Jul 11;8(7):e020943 [FREE Full text] [doi: 10.1136/bmjopen-2017-020943] [Medline: 29997138]
29. Xu D, Pan J, Dai X, Hu M, Cai Y, He H, et al. Comparing quality of primary healthcare between public and private providers in China: study protocol of a cross-sectional study using unannounced standardised patients in seven provinces of China. BMJ Open 2021 Jan 12;11(1):e040792 [FREE Full text] [doi: 10.1136/bmjopen-2020-040792] [Medline: 33436467]
30. Xu D, Hu M, He W, Liao J, Cai Y, Sylvia S, et al. Assessing the quality of primary healthcare in seven Chinese provinces with unannounced standardised patients: protocol of a cross-sectional survey. BMJ Open 2019 Feb 13;9(2):e023997 [FREE Full text] [doi: 10.1136/bmjopen-2018-023997] [Medline: 30765399]
31. Zhou M, Wang H, Zhu J, Chen W, Wang L, Liu S, et al. Cause-specific mortality for 240 causes in China during 1990-2013: a systematic subnational analysis for the Global Burden of Disease Study 2013. Lancet 2016 Jan 16;387(10015):251-272. [doi: 10.1016/S0140-6736(15)00551-6] [Medline: 26510778]
32. McCabe D, Langer KG, Borod JC, Bender HA. Practice Effects. In: Kreutzer JS, DeLuca J, Caplan B, editors. Encyclopedia of Clinical Neuropsychology. New York, NY: Springer; 2011.
33. Tamblyn RM. Use of standardized patients in the assessment of medical practice. CMAJ 1998 Jan 27;158(2):205-207 [FREE Full text] [Medline: 9469142]
34. Xu DR, Cai Y, Wang X, Chen Y, Gong W, Liao J, et al. Improving data surveillance resilience beyond COVID-19: experiences of primary heAlth Care quAlity Cohort In ChinA (ACACIA) using unannounced standardized patients. Am J Public Health 2022 Jun;112(6):913-922. [doi: 10.2105/AJPH.2022.306779] [Medline: 35483014]
35. Gehanno JF, Rollin L, Le Jean T, Louvel A, Darmoni S, Shaw W. Precision and recall of search strategies for identifying studies on return-to-work in Medline. J Occup Rehabil 2009 Sep;19(3):223-230. [doi: 10.1007/s10926-009-9177-0] [Medline: 19381789]
36. Chon S, Hilgers S, Timmermann F, Dratsch T, Plum PS, Berlth F, et al. Web-based immersive patient simulator as a curricular tool for objective structured clinical examination preparation in surgery: development and evaluation. JMIR Serious Games 2018 Jul 04;6(3):e10693 [FREE Full text] [doi: 10.2196/10693] [Medline: 29973333]
37. Olgers TJ, Bij de Weg AA, Ter Maaten JC. Serious games for improving technical skills in medicine: scoping review. JMIR Serious Games 2021 Jan 25;9(1):e24093 [FREE Full text] [doi: 10.2196/24093] [Medline: 33492234]
38. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med 2018 Sep;18(3):91-93 [FREE Full text] [doi: 10.1016/j.tjem.2018.08.001] [Medline: 30191186]
39. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
40. Jakobsson U, Westergren A. Statistical methods for assessing agreement for ordinal data. Scand J Caring Sci 2005 Dec;19(4):427-431. [doi: 10.1111/j.1471-6712.2005.00368.x] [Medline: 16324069]

41. Li X, Lu J, Hu S, Cheng KK, De Maeseneer J, Meng Q, et al. The primary health-care system in China. Lancet 2017 Dec 09;390(10112):2584-2594. [doi: 10.1016/S0140-6736(17)33109-4] [Medline: 29231837]
42. Sylvia S, Xue H, Zhou C, Shi Y, Yi H, Zhou H, et al. Tuberculosis detection and the challenges of integrated care in rural China: A cross-sectional standardized patient study. PLoS Med 2017 Oct;14(10):e1002405 [FREE Full text] [doi: 10.1371/journal.pmed.1002405] [Medline: 29040263]
43. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
44. Das J, Kwan A, Daniels B, Satyanarayana S, Subbaraman R, Bergkvist S, et al. Use of standardised patients to assess quality of tuberculosis care: a pilot, cross-sectional study. Lancet Infect Dis 2015 Nov;15(11):1305-1313 [FREE Full text] [doi: 10.1016/S1473-3099(15)00077-8] [Medline: 26268690]
45. Mohanan M, Vera-Hernández M, Das V, Giardili S, Goldhaber-Fiebert JD, Rabin TL, et al. The know-do gap in quality of health care for childhood diarrhea and pneumonia in rural India. JAMA Pediatr 2015 Apr 01;169(4):349-357 [FREE Full text] [doi: 10.1001/jamapediatrics.2014.3445] [Medline: 25686357]
46. Weiner SJ, Schwartz A. Directly observed care: can unannounced standardized patients address a gap in performance measurement? J Gen Intern Med 2014 Aug;29(8):1183-1187 [FREE Full text] [doi: 10.1007/s11606-014-2860-7] [Medline: 24756945]
47. Muse K, McManus F. A systematic review of methods for assessing competence in cognitive-behavioural therapy. Clin Psychol Rev 2013 Apr;33(3):484-499. [doi: 10.1016/j.cpr.2013.01.010] [Medline: 23454222]
48. Chinese Medical Association, Chinese Medical Journals Publishing House, Chinese Society of Gastroenterology, Chinese Society of General Practice, Editorial Board of Chinese Journal of General Practitioners of Chinese Medical Association, Expert Group of Guidelines for Primary Care of Gastrointestinal Disease. Guidelines for primary care in chronic gastritis (2019). Chinese Journal of General Practitioners 2020 Sep 04;19(09):768-775. [doi: 10.3760/cma.j.cn114798-20200706-00785]
49. Phichitchaisopa N, Naenna T. Factors affecting the adoption of healthcare information technology. EXCLI J 2013;12:413-436 [FREE Full text] [Medline: 26417235]
50. Thompson J, White S, Chapman S. Interactive clinical avatar use in pharmacist preregistration training: design and review. J Med Internet Res 2020 Nov 06;22(11):e17146 [FREE Full text] [doi: 10.2196/17146] [Medline: 33155983]
51. Watari T, Tokuda Y, Owada M, Onigata K. The utility of virtual patient simulations for clinical reasoning education. Int J Environ Res Public Health 2020 Jul 24;17(15):5325 [FREE Full text] [doi: 10.3390/ijerph17155325] [Medline: 32722097]
52. Ben-Assuli O, Kumar N, Arazy O, Shabtai I. The use of analytic hierarchy process for measuring the complexity of medical diagnosis. Health Informatics J 2020 Mar;26(1):218-232 [FREE Full text] [doi: 10.1177/1460458218824708] [Medline: 30672359]
53. Shorey S, Ang E, Ng ED, Yap J, Lau LST, Chui CK. Communication skills training using virtual reality: A descriptive qualitative study. Nurse Educ Today 2020 Nov;94:104592. [doi: 10.1016/j.nedt.2020.104592] [Medline: 32942248]
54. Benda NC, Kellogg KM, Hoffman DJ, Fairbanks RJ, Auguste T. Lessons learned from an evaluation of serious gaming as an alternative to mannequin-based simulation technology: randomized controlled trial. JMIR Serious Games 2020 Sep 28;8(3):e21123 [FREE Full text] [doi: 10.2196/21123] [Medline: 32985993]
55. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA 1999 Oct 20;282(15):1458-1465. [doi: 10.1001/jama.282.15.1458] [Medline: 10535437]
56. Paksaite P, Crosskey J, Sula E, West C, Watson M. A systematic review using the Theoretical Domains Framework to identify barriers and facilitators to the adoption of prescribing guidelines. Int J Pharm Pract 2021 Feb 10;29(1):3-11. [doi: 10.1111/ijpp.12654] [Medline: 32656929]
57. Almazrou SH, Alfaifi SI, Alfaifi SH, Hakami LE, Al-Aqeel SA. Barriers to and facilitators of adherence to clinical practice guidelines in the Middle East and North Africa region: a systematic review. Healthcare (Basel) 2020 Dec 15;8(4):564 [FREE Full text] [doi: 10.3390/healthcare8040564] [Medline: 33333843]

Abbreviations

CCC: concordance correlation coefficient
PHC: primary health care
USP: unannounced standardized patient
VP: virtual patient

Edited by T Leung; submitted 05.06.22; peer-reviewed by H Zheng, M Jones; comments to author 06.09.22; revised version received 27.09.22; accepted 29.10.22; published 02.12.22

Please cite as:
Zeng M, Cai Y, Cao J, He Q, Wang X, Lu Y, Liang H, Xu D, Liao J
The Agreement Between Virtual Patient and Unannounced Standardized Patient Assessments in Evaluating Primary Health Care Quality: Multicenter, Cross-sectional Pilot Study in 7 Provinces of China
J Med Internet Res 2022;24(12):e40082
URL: https://www.jmir.org/2022/12/e40082
doi: 10.2196/40082
PMID: