BMJ Open

Using smartphone-based virtual patients to assess the quality of primary healthcare in rural China: protocol for a prospective multicentre study

Jing Liao,[1] Yaolong Chen,[2] Yiyuan Cai,[3] Nan Zhan,[4] Sean Sylvia,[5] Kara Hanson,[6] Hong Wang,[7] Judith N Wasserheit,[8] Wenjie Gong,[9] Zhongliang Zhou,[10] Jay Pan,[11] Xiaohui Wang,[12] Chengxiang Tang,[13] Wei Zhou,[14] Dong Xu[1]

To cite: Liao J, Chen Y, Cai Y, *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;**8**:e020943. doi:10.1136/bmjopen-2017-020943

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-020943).

Received 8 December 2017
Revised 16 May 2018
Accepted 7 June 2018

For numbered affiliations see end of article.

Correspondence to
Dr Dong Xu;
xudong5@mail.sysu.edu.cn

ABSTRACT

Introduction Valid and low-cost quality assessment tools examining care quality are not readily available. The unannounced standardised patient (USP), the gold standard for assessing quality, is costly to implement while the validity of clinical vignettes, as a low-cost alternative, has been challenged. Computerised virtual patients (VPs) create high-fidelity and interactive simulations of doctor-patient encounters which can be easily implemented via smartphone at low marginal cost. Our study aims to develop and validate smartphone-based VP as a quality assessment tool for primary care, compared with USP.

Methods and analysis The study will be implemented in primary health centres (PHCs) in rural areas of seven Chinese provinces, and physicians practicing at township health centres and village clinics will be our study population. The development of VPs involves three steps: (1) identifying 10 VP cases that can best represent rural PHCs' work, (2) designing each case by a case-specific development team and (3) developing corresponding quality scoring criteria. After being externally reviewed for content validity, these VP cases will be implemented on a smartphone-based platform and will be tested for feasibility and face validity. This smartphone-based VP tool will then be validated for its criterion validity against USP and its reliability (ie, internal consistency and stability), with 1260 VP/USP-clinician encounters across the seven study provinces for all 10 VP cases.

Ethics and dissemination Sun Yat-sen University: No. 2017-007. Study findings will be published and tools developed will be freely available to low-income and middle-income countries for research purposes.

Strengths and limitations of this study

- ► Developing and validating smartphone-based virtual patient (VP) as a quality assessment tool for research and routine use in rural primary healthcare centres.
- ► Following an evidence-based approach to develop VP cases and scoring criteria.
- ► Systematically validating the VP assessment tool via a cross-national multicentre study.
- ► The extent to which the VP assessment will reflect practitioners' real clinical practice needs to be verified.

INTRODUCTION

Universal health coverage (UHC) is a paramount goal of health system development for countries at all income levels.[1] The achievement of UHC is not possible without primary healthcare services,[1] which ensure integrated care close to the population they serve and link to the health-related sustainable development goals.[2] However, service coverage alone cannot improve health outcomes if the quality of care is poor. Despite efforts devoted to improving healthcare services, there is a lack of scientific evidence on the quality of primary healthcare in resource-poor settings, particularly of low-income and middle-income countries (LMICs).[3–6]

This scarcity of evidence may partially result from the limited availability of valid, low cost and easy-to-implement quality assessment tools.[7] As defined by Donabedian's framework, healthcare quality can be evaluated by the *structure* of care (eg, staff, equipment), the *process* of care delivery (eg, doctor-patient interactions) and health *outcomes* (eg, death or complications).[8] Increasingly, process measures are being used, because of their advantages in terms of frequent and timely evaluation and the usefulness in improving practice.[9 10] The 'gold standard' of assessing process is the unannounced standardised patient (USP), namely a trained actor who simulates the symptoms, signs and emotions of a real patient in a standardised fashion and presents himself or herself unannounced to clinics to assess care quality.[11] USP can reduce recall bias better than patient exit interviews, minimise the Hawthorne effect

that inevitably occurs in direct observation and allow for comparisons between users as case-mix and patient-mix are controlled.[3 9 11] Nonetheless, the USP can only portray a limited number of conditions without obvious physiological symptoms and risk of invasive examinations. Also, training and implementation of USP can require substantial personnel and resources, making USP impractical for large-scale and routine quality assessment.[12 13]

As an alternative, clinical vignettes or case simulations have been widely used as a low-cost and convenient method for assessing care quality.[9 14] Vignettes have been implemented in a paper-and-pencil form,[9] presented by an enumerator[5] and streamlined by a computer.[14] Evidence of the validity of vignettes in assessing the quality of patient care is mixed. Some studies showed that vignettes reflect clinicians' competency (know-how) rather than their actual behaviours and can lead to overestimation of clinical performance.[9 15] By contrast, other studies found that vignette-based results, particularly those streamlined by computer, are quite close to the USP-based assessment.[14 16] The enumerator-administered vignette is similar to the announced standardised patient and thus is expensive and difficult to implement.[5] A computerised vignette can be interactive and can more realistically represent the complexity of a clinical encounter.[14] As a further improvement on computerised vignettes,[14] smartphone virtual patients (VPs) create high-fidelity, visualised and interactive simulations that replicate clinical complexity and can be easily implemented at a low marginal cost.[17] Although VPs cannot remove the Hawthorne effect, their advanced features may reduce the measurement gap between competency and actual practice.[10 14] While VPs have been used in medical education to train and test clinical skills such as clinical reasoning, diagnosis and therapeutic decisions,[18] their relative validity as a measure of quality of care has yet to be studied. Strengths and limitations of the abovementioned three methods are compared in table 1.

In the present study, we propose to adapt smartphone-based VP for medical education as a quality-of-care assessment tool, given its advantages in (1) standardisation (VPs are highly standardised, ensuring consistent assessments across users), (2) *flexibility* (assessments can be delivered by smartphones for multiple users at any time, anywhere, providing data connectivity is available), (3) *scalability* (VPs can be modified to demonstrate and assess almost any clinical conditions with low marginal cost) and (4) *training* (VPs can also be used as a training tool to improve healthcare quality and thus to address the 'so what' question after quality assessment). These characteristics may especially benefit quality assessment and improvement in rural primary care settings, where communities are geographically scattered and difficult to reach and manage.

Therefore, our study aims to develop and validate *smartphone-based VPs* against USPs as a quality assessment tool that can be used both for research purposes and for routine evaluation of quality of primary healthcare provided by primary health centres (PHCs) in rural areas. To maximise its validity,[14] we will systematically construct high-fidelity VP cases to reflect clinical complexity in rural PHC contexts with real-time patient-doctor interactions and temporal constraints and use evidence-based quality scoring criteria; additionally, we will make the VP-based test anonymous to minimise the Hawthorne effect. The initial phase of the study will mainly focus on rural China, while the ultimate goal is to develop and validate tools that can have a broad application in other LMICs.

METHODS AND ANALYSIS

Study setting

The validation study will be implemented in the outpatient setting of rural PHCs (ie, township health centres and village clinics) in seven Chinese provinces (Guizhou,

Table 1 Strengths and limitations of quality assessment measures

Process measure of quality	Strengths	Limitations	Assessment level
Unannounced standardised patient	Standardised, controlling for case-mix and patient- mix; Unannounced, no Hawthorne effect (if not be detected); Reduced recall bias	Expensive Limited medical conditions First-visit bias Selection bias if informed consent is required	*Action* (Do) *Gold standard*
Clinical vignettes	Cost-effective; Suitable for large scale and cross-system comparison; Covering all illnesses; Accounting for case-mix and patient-mix.	Hawthorne effect Selection bias	*Competence* (Know how) *Overestimating* care quality (best answers) or similar
Virtual patients	Interactive Real-time response and automatic record Highly standardised Scalability and low marginal cost Efficient delivery: anytime, anywhere Suitable for large scale study	Hawthorne effect Selection bias High cost in initial development	*Performance* (Show how) *?*

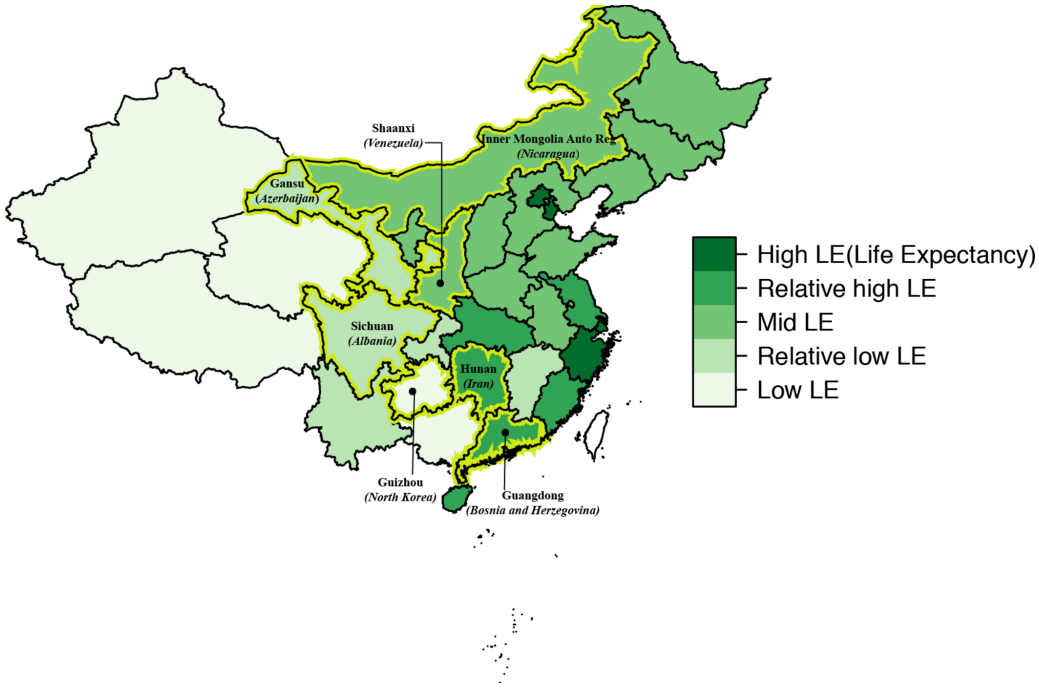

Figure 1 Seven sample provinces in China referencing countries of equivalent life expectancy in parentheses.

Sichuan, Gansu, Inner Mongolia, Shaanxi, Hunan and Guangdong). We are selecting these provinces to reflect the five strata of low-to-high life-expectancies and various burden-of-disease patterns in China[19] and to contrast geographic regions with diverse ethnic composition, including southwest mountainous regions, the northern plateau, the middle inland region and southeast coastal areas (figure 1). Our study targets township health centres and village clinics because they provide the majority of primary healthcare in rural China.[20 21] At township health centres, primary healthcare is delivered by a workforce including licensed/unlicensed physicians, licensed/unlicensed assistant physicians and registered nurses, while at village clinics, services are mainly delivered by one full-time or part-time 'village doctor' who is a clinician with rudimentary medical training.[20 22 23] The outpatient setting is chosen due to the small number of inpatient cases in township health centres and village clinics. Study recruitment is expected to start from June 2018.

VP case development

VP case selection

We intend to select 10 cases that together can represent the work of rural PHCs. The selection of the VP cases will be based on the following criteria: (1) high frequency of clinical encounters in the primary care settings in rural areas and/or (2) association with significant disease burden; (3) representation of the major areas of work of PHCs in rural China overall (eg, public health service delivery, chronic disease management, infectious disease control, health education and patient-centred care) and (4) suitability for the USP methodology (eg, no obvious physiological signs, low risk for invasive tests) for the sake of criterion validation in the current study. A case selection committee will comprise a range of stakeholders, including physicians, public health practitioners, policy-makers and members of the research team. Based on the literature review, the research team will prepare a shortlist of the 30 most frequently seen conditions in township health centres and village clinics reported by either community dwellers[24] or rural PHC clinicians (online supplementary appendix 1) from which the committee will select.

VP case design

The 10 selected VP cases will then be constructed individually by 10 case-specific development teams (figure 2). These teams consist of one *condition expert* from the relevant specialty of a tertiary teaching hospital who will be responsible for drafting the VP case; an *evidence-synthesis group* involving epidemiologists and evidence-based researchers who will search and synthesise evidence about the selected condition for the condition expert to work on; a *clinical consensus group* which consists of several condition-related clinical experts who will review the corresponding case from a scientific perspective; an overall all-condition shared *context-expert panel*, which includes clinicians and health managers from community health centres, township health centres and village clinics, who will review the contextual appropriateness of the cases for the rural PHC setting and a *case coordinator* who will coordinate development of each case.

Each VP case will be structured into five domains—medical history, physical examination, laboratory and imaging studies, diagnosis and management and treatment plan—to simulate real-life clinical scenarios.[11 18] The structured VP cases will permit the examinee's performance in each domain to be evaluated and for performance scores

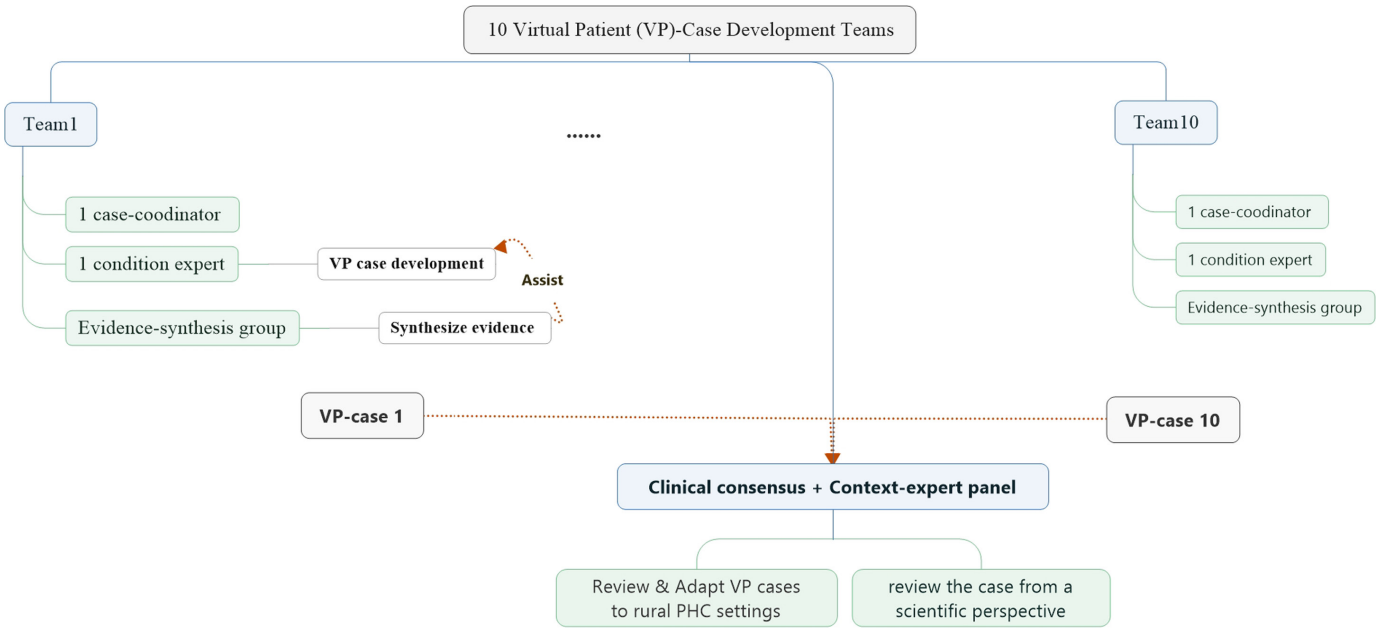

Figure 2 Virtual patient case development team role and responsibilities.

to be aggregated across conditions. In addition to these five condition-related domains, another practice contextual adjustment will be built into each case to consider medical resource constraints in rural practices (eg, availability of basic medical equipment and medicines).

Scoring criteria

Care quality scoring criteria will be developed for each VP case. These criteria include *process quality*, the *accuracy of diagnosis and* the *appropriateness of the treatment and management plan*.[3 4] Process quality will be evaluated in reference to a clinical process checklist (to be detailed later) including all necessary questions that should be asked and physical examinations that should be performed by clinicians, together with redundant or even potentially harmful practices. Diagnoses will be rated as correct, partially correct or incorrect based on predetermined standards. The treatment and management plan will be considered appropriate if the clinician prescribes any of the correct medications or refers the patient to a higher-level physician depending on the VP case.

In addition, *cost of care* and *time-spent per encounter* will also be recorded. Patient costs will cover medication fees and clinic fees charged per case. In order to link clinician reaction time to each domain and to impose the temporal constraints seen in real clinical practices, the entire clinician-VP interaction process will be timed. This will include time spent on taking history, conducting physical examinations, prescribing drugs and treatments and any interruptions.

A systematic evidence-based approach will be adopted to developing the scoring checklist for the treatment and management plan (online supplementary appendix 2). Briefly, the evidence-synthesis group will systematically search and extract condition-specific checklist items and standards from clinical guidelines, reputable textbooks and systematic reviews, among others, in that order. The quality of the evidence will then be rated by the Appraisal of Guidelines for Research & Evaluation II (AGREE II)[25] or the revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) according to its type.[26] Afterwards, the clinical consensus group and the contextual-expert panel will review and revise initial standards using a Delphi process.[27]

VP case external review

To validate the content of VP cases, an independent expert panel of physicians, general practitioners and rural PHC clinicians who otherwise are not involved in the study, will be convened to review the cases for content accuracy and appropriateness. The content validation involves qualitative and quantitative phases. In the qualitative phase, the expert panel will be required to evaluate the cases with respect to the following: overarching assessment goal, representativeness of the goal and test items to each domain, the logical relationship of the content tested and the appropriate wording, grammar, understandability and relevance to the rural PHC context. The panel will also record their suggestions, if any, next to each item. Modified VP cases will then be given back to the expert panel for quantitative evaluation. They will be asked to assess the cases for simplicity and clarity as well as necessity and relevance to the assessment, using a four-point Likert scale ranging from 1 (the lowest) to 4 (the highest). The content validity index will be computed for each domain and for the entirety of VP cases.[28]

Technical implementation

Revised VP cases will be implemented on *CureFUN*, an existing smartphone-based training platform using VPs with special customisations and set-up to suit the assessment purpose. A live demonstration of a simplified VP

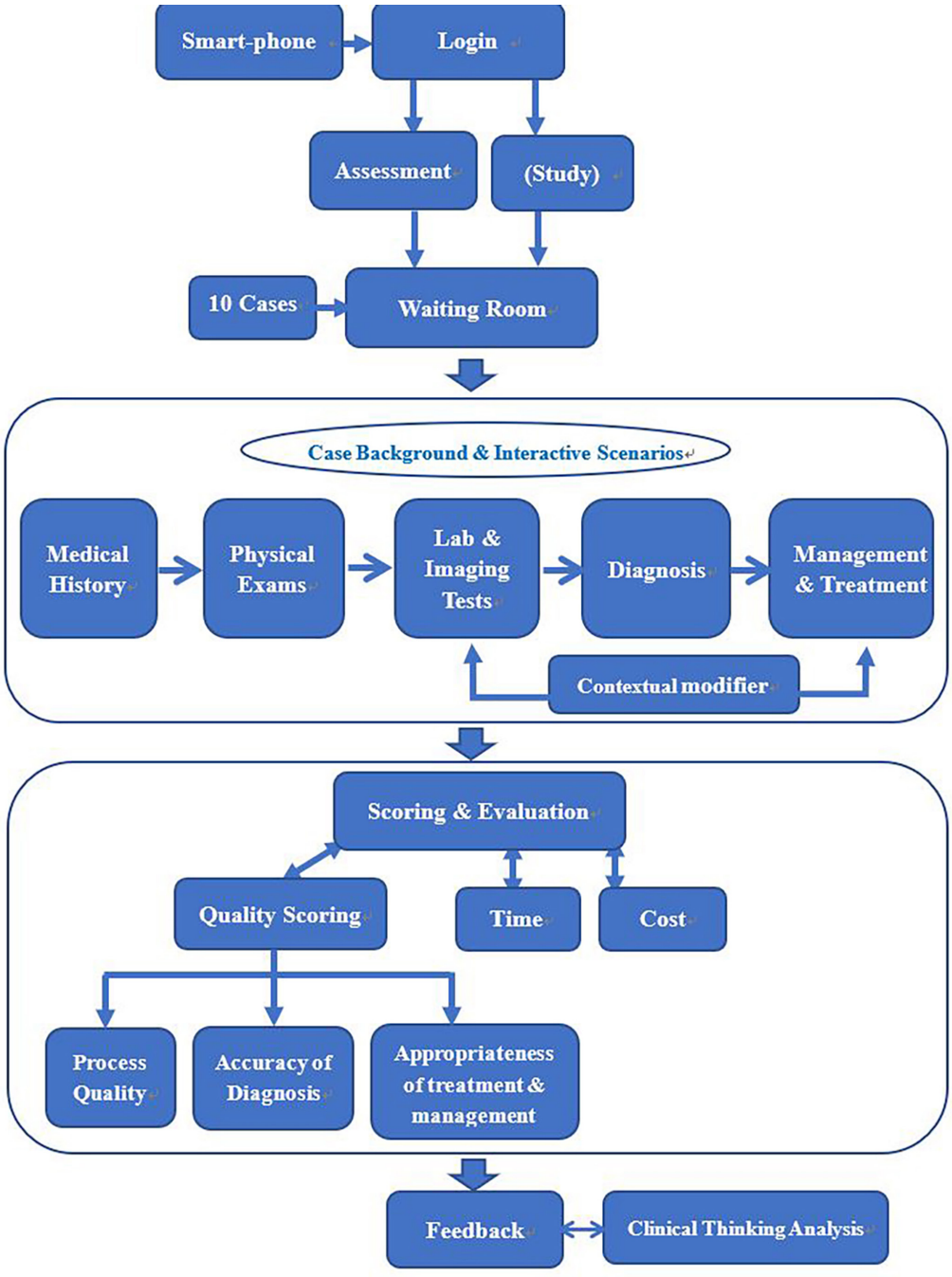

Figure 3 Main components of smartphone-based virtual patient programme.

can be accessed from http://www.curefun.com/zhiqu_front/www/experience/experience.html#/caseList and is also illustrated in online supplementary appendix 3. The smartphone-based VP assessment tool will present interactive clinical scenarios and will automatically record each examinee's diagnosis pathway and grade it against the scoring criteria (figure 3).

Feasibility study

Before a full-scale validation study, a feasibility study with 30% of the validation study sample (see study sample section) will be conducted to test the VP assessment tool's usability, accessibility and stability, particularly in remote village clinics with weak phone connectivity. Selected clinicians will be instructed to individually attempt two random VP cases within a given time, using their own smartphone devices from their workplace. Clinicians without a smartphone will be given a temporary device on which the customised *CureFUN* applications will be preinstalled. Clinicians' willingness to participate and adherence to the VP-based tests (eg, percentage completing VP cases, score of the assessment and number of attempts made at each case per person) will be automatically recorded. On completion of the cases, participants will be asked to fill in a five-point Likert-scale questionnaire regarding their

subjective attitude towards the simulator VP experience (with 1 being the most negative response and 5 being the most positive), regarding ease of use, their experience of the assessment process and outcome, realism, device competence, accessibility and other general comments. These results will be used to determine the face validity of the VP cases, with scores calculated by multiplying frequency (%) by positive evaluations (3–5); and scores no less than 1.5 are considered acceptable.[29]

Validation of VP as a quality assessment tool

Study design

The prospective validation study is a nationwide multicentre study with two main purposes: (1) to assess the criterion validity of the VP-tool in assessing the quality of primary healthcare, by analysing its measurement concordance against the standard USP measure and (2) to test the reliability of the VP tool, by examining its internal consistency and the stability of repeated VP assessments on the same subjects.

Study sample

From each of our seven sample provinces, two counties will be selected with sufficient variations in socioeconomic conditions, demographics and disease burdens between them while also approximating the provincial condition in general. Within each county, the government registry of all township health centres and village clinics will serve as our sampling frame, which will include (1) licensed practicing physicians, (2) clinicians who have not been licensed but are providing clinical services under the supervision of licensed physicians at township health centres as well as (3) full-time or part-time village clinicians. Clinicians visiting on a temporary basis (often senior clinicians sent by higher level medical institutions to support the development of township health centres), nurses and allied health workers without prescription privileges will be excluded.

The sample size calculation is based on individual VP/USP-clinician encounter and ensures sufficient power to detect variations at individual case level per county. For village clinics, one VP/USP case will be examined at a time to minimise the detection of USPs. Assuming a 5% type I error and 80% power, to determine whether a moderate concordance correlation coefficient[30] of 0.90[31] between VP and USP differs from zero, seven paired VP/USP-clinician encounters will be required for each of the 10 cases per county. As a stratified sampling strategy will be deployed that first samples townships and then villages from each township, sample size calculations need to take into account the design effect. Assuming an intraclass correlation of 0.05 and six village clinics per township, then nine paired VP/USP-clinician encounters is needed. These nine paired VP/USP-clinician encounters will be assigned to three township health centres and six village clinics using probability proportional to size. These nine paired VP/USP-clinician encounters will be assigned to township health centres and village clinics based on the ratio of the total number of clinicians at township health centres to the total number of village clinicians for each county. There are 1260 VP/USP-clinician encounters across our 7 study provinces for all 10 VP cases. Figure 4 shows the sampling process and study flow for one VP case using Guizhou Province (Danzhai County) as an example.

Criterion validity

Criterion validity[32] of the VP to assess quality of care will be evaluated primarily by its measurement concordance against the USP measure as the recognised gold standard[33] for assessing quality of care in practice. The USPs will be developed in a related study, sharing the development teams for VP and a similar development process. The method of fielding USPs in rural China will follow a similar approach to those of the previous USP study in rural China.[3] Identical quality scoring criteria, described above, will be applied to scores. Each selected clinician will first see a USP (to avoid the practice effect due to the USP's unannounced feature) and then complete a smartphone-based VP assessment of the same condition. The clinician to be assessed will be randomly selected onsite by the USP from any on-duty clinicians on the day of the USP visit to the sampled township health centre and village clinics. This situation would especially apply to township health centres, as most village clinics have only one clinician (note: Chinese patients normally see their primary care clinicians as a walk-in patient and appointments are seldom needed). To record USP-clinician interactions, USPs will complete checklists immediately after their visit and retain their prescription and the fee charge slips provided by the clinician. A week after the USP clinic visit, clinicians will be assigned a smartphone-based VP assessment, which will consist of an initial demonstration VP case to allow the clinician to familiarise themselves with how the system operates and then the test VP case of the same USP condition. The VP-clinician interactions, drugs dispensed and fees charged will all be recorded automatically by the online assessment system.

The concordance of the two USP and VP assessments will then be analysed by Lin's concordance correlation coefficient (r_c)[30] for continuous *process quality* scores, *fees charged* (yuan) and *time spent* (min) and the Kappa statistic[34] for dichotomous *diagnoses* and *treatment & management* measures. r_c evaluates how close pairs of observation fell on a 45° line (the perfect concordance line) through the origin in addition to their correlation. Kappa measures agreement in assessment beyond what is expected by chance alone. In addition, for continuous measures, a Bland-Altman plot will also be used to visualise the concordance.[35 36] For dichotomous measures, we will analyse their sensitivity (ie, strength to detect correct diagnosis, treatment plan, among others) and specificity (ie, strength to detect incorrect diagnosis, treatment plan, among others) using USP as the reference.

Reliability

To establish *test-retest reliability*, clinicians previously being assessed by VPs will be instructed to retake the same VP tests

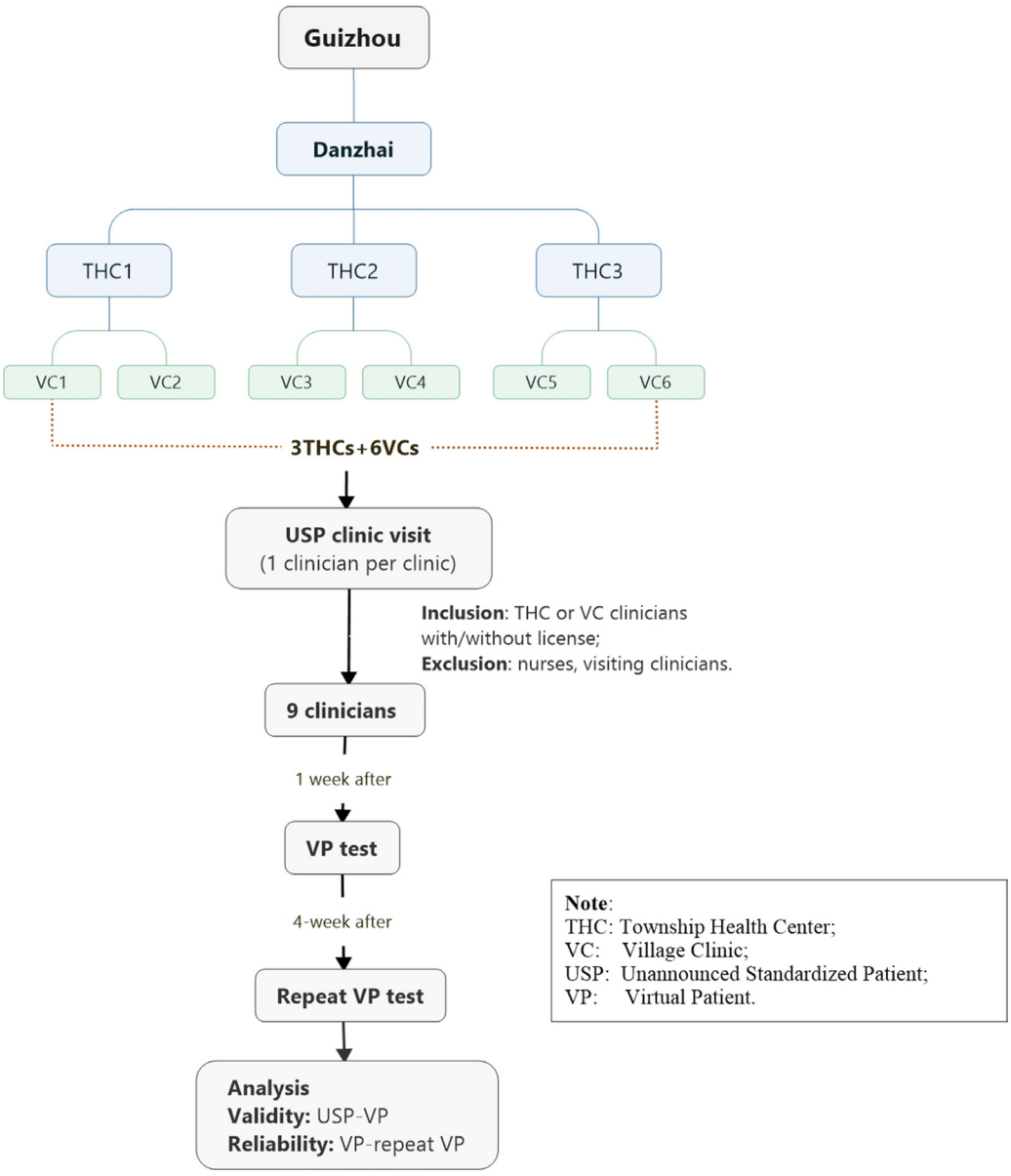

Figure 4 Sampling and study process for one VP case in Danzhai County, Guizhou Province.

4 weeks after their last assessment. The second VP test is set 1 month later than the first to reduce the practice effect,[37] assuming the clinician's general medical knowledge remains constant.[32] The concordance of the two repeated tests indicates the stability of the VP assessment tool.[32] Similar concordance measures (ie, r_c for continuous and Kappa for dichotomous measures) as described above will be used. *Internal consistency*, the intercorrelation of scores for process quality indicators, will be computed by Cronbach's alpha coefficients (α),[38] with $\alpha > 0.7$ representing acceptable reliability.[39] Table 2 summarises the validity, reliability and feasibility measures that will be examined in our study.

Patient and public involvement

We will seek feedback from clinicians and patient representatives in the feasibility study and use their feedback to refine the VP cases. Our USPs will be lay people trained to portray patients and assess care quality based on their interactions with clinicians. Our scoring criteria thus are also patient-centred. Furthermore, all participants will be acknowledged for their involvement in the study and will be provided with a final summary report of the study outcomes and will have free access to the VP training website. All published results will be publicly available.

ETHICS AND DISSEMINATION

Informed consent will be obtained from all clinicians participating in the VP tests. However, to reduce participation bias due to self-selection,[40] our IRB has approved the implementation of USP without prior informed consent from the individual participants, on the condition that involved clinicians will be fully deidentified and all analyses will only be conducted at the population level.[40] Study data will be securely stored and only deidentified information

Table 2 Main validation domains of the study

Domain	Indicator	Data collection		Statistical analysis
		Phase	Method	
Content validity	Content validity index	VP case review	Evaluations by an expert panel after reviewing VP cases, measured by a 4-point Likert scale (1=lowest, 4=highest).	CVI for VP case and for specific VP domain will be computed, where CVI=number of raters giving a rating of 3 or four divided by the total number of raters.
Feasibility	Willingness to participate; Adherence rate	Feasibility study	The subsample of clinicians' interactions with the two VP cases will be recorded by the online assessment	Willingness to Participate=clinicians taking the VP tests divided by the percentage of clinician selected Adherence rate=clinicians completed two VP cases divided by the percentage of clinicians taking VP tests
Face validity	Satisfying score		Clinicians' subjective attitude towards the VP test experience measured by a 5-point Likert scale (1=most negative, 5=most positive).	Satisfying score for VP case and for specific aspects (eg, usability, accessibility, etc) will be computed, where satisfying score=frequency multiply by positive evaluations (3–5) and scores≥1.5 are considered acceptable.
Criterion validity	Concordance correlation coefficient (r_c); Kappa statistic	Validation study	The same clinician receives a USP visit and a VP test for a matching condition. The USP-clinician interaction is evaluated by the USP using the checklist, including fees and time per visit; while VP-clinician interaction is graded by the system.	The concordance of VP-test scores against USP-test score (gold standard) or two-repeated VP-tests will be examined by r_c for continuous process quality scores, fees charged (yuan) and time spent (min) and Kappa for dichotomous diagnoses and treatment and management measures.
Test-retest reliability			Repeat VP-tests on the same clinician in a month	
Internal consistency	Cronbach's alpha coefficient (α)		VP-test scores on a single occasion	Intercorrelation of scores for process quality indicators with α>0.7 is acceptable.

CVI, content validity index; VP, virtual patient.

will be used for analysis. We will seek to publish study findings in peer-reviewed journals and produce reports to inform health authorities. The tools and technology developed in this study will be freely available to other LMICs for research purposes.

DISCUSSION

To the best of our knowledge, this is the first study validating VP as a quality assessment tool in rural primary healthcare centres. This study follows an evidence-based approach to develop VP cases and scoring criteria, implements them on a widely accessible platform (ie, a smartphone) and systematically validates the VP assessment tool via a cross-national multicentre study representing rural PHCs over a wide range of geographic areas with distinct life expectancies and economic development levels. The VP assessment tool's accessibility, flexibility and scalability give it good potential to be easily adapted to other LMICs.

VP has mainly been used in medical education to train and test critical thinking,[18 41 42] and until recently few studies have applied the method in a practice setting to influence health provider behaviour and improve care quality.[43 44] As an extension, we propose to validate VP as a quality assessment tool delivered via widely accessible smartphones. Nevertheless, it is to be noted that given its simulated nature, the VP-test theoretically may never completely bridge the 'know-do' gap. The validation study is thus essential to quantify the concordance/discordance between VP-based and USP-based quality assessments. Our study will generate firsthand empirical evidence contributing to the understanding of the 'know-do gap'[5 45] and shed light on circumstances that cannot be tested by USPs.

A limitation of the study, however, is that, in order to test the validity of VP against USP as the reference standard, we restrict the selection of VP cases to those that can be simulated by USP. This conservative first step will nevertheless allow us to examine the extent to which VP can reflect care quality, and a follow-up study will then explore the full potential of the VP in assessing quality of care. Further, the two purposely selected counties for each province may not represent the provincial conditions entirely, although we will make every effort to consider provincial representation when selecting counties. Third, while the validation study is exclusively conducted on PHCs in rural China, the

extent to which the VP assessment tool can be transported to other LMICs remains to be evaluated. Nonetheless, by implementing the study in a diverse set of Chinese provinces may improve the generalisability of our study considering the comparable life expectancies of LMICs and these provinces.

Author affiliations
[1]Sun Yat-sen Global Health Institute, School of Public Health and Institute of State Governance, Sun Yat-sen University
[2]Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou, China
[3]School of Public Health, Guizhou Medical University, Guiyang, China
[4]Department of Health Management, School of Health Management, Inner Mongolia Medical University, Hohhot, China
[5]Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[6]Department of Global Health and Development, Faculty of Public Health and Policy, London School of Hygiene & Tropical Medicine, London, UK
[7]Health Economics, Financing & Systems, Bill & Melinda Gates Foundation, Seattle, Washington, USA
[8]Departments of Global Health, Medicine, and Epidemiology, Schools of Medicine and Public Health, University of Washington, Seattle, Washington, USA
[9]Xiangya School of Public Health, Central South University, Changsha, China
[10]School of Public Policy and Administration, Xi'an Jiaotong University, Xi'an, China
[11]West China School of Public Health, Sichuan University, Chengdu, China
[12]School of Public Health, Lanzhou University, Lanzhou, China
[13]School of Public Administration, Guangzhou University, Guangzhou, China
[14]Hospital Administration Institute, Xiangya Hospital, Central South University, Changsha, China

Acknowledgements The study will be led by the Sun Yat-sen Global Health Institute of Sun Yat-sen University with a consortium of researchers from seven other Chinese universities, including Central South University, Guangzhou University, Guizhou Medical University, Inner Mongolia Medical University, Lanzhou University, Sichuan University and Xi'an Jiaotong University. We thank all the students from these universities who have contributed to our project, especially Wenjun He who produced figure 1, Jianjian Wang who assisted with the evidence collection of the Scoring Criteria and Ash Harris who contributed to the VP computerisation.

Contributors All authors contributed to the conceptualisation and design of the study. DX, JL and YC conceived the initial study design, analytical methods and composition of the team. JL was responsible for the study concept, initial draft and revisions. DX was responsible for the study concept and revising the draft. YC and XW were responsible for the development of the scoring criteria. SS, KH, HW, JNW, ZZ, NZ, WG, JP, CT and WZ provided critical review and revision to the study design. All authors read and approved the final revision.

Funding The study is supported by the China Medical Board through its Health Policy and Systems Sciences Open Competition grant 'Quality in primary health care: using unannounced standardised patients' (grant No.: CMB16-260, XD, PI).

Competing interests The development of the VP assessment tool is a joint project of the Sun Yat-sen University Global Health Institute (SGHI) (representing the seven universities in China) and CureFUN. However, the VP cases will be independently developed and the validation studies will be rigorously conducted by the research team from SGHI and the seven universities, whereas CureFun will technically implement the cases on smartphones and have no influence over the study design and analysis.

Patient consent Obtained.

Ethics approval The Institutional Review Board of the School of Public Health (IRB), Sun Yat-sen University (No. 2017-007).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement We will seek peer-reviewed publications for study findings and produce reports to inform health authorities. The tools and technology developed in this study will be freely available to other LMICs for research purposes.

REFERENCES

1. Kieny MP, Evans DB. Universal health coverage. *East Mediterr Health J* 2013;19:305–6.
2. United Nations- Sustainable Development Knowledge Platform. *Transforming our world: the 2030 Agenda for Sustainable Development*, 2015.
3. Sylvia S, Shi Y, Xue H, *et al*. Survey using incognito standardized patients shows poor quality care in China's rural clinics. *Health Policy Plan* 2015;30:322–33.
4. Das J, Holla A, Das V, *et al*. In urban and rural India, a standardized patient study showed low levels of provider training and huge quality gaps. *Health Aff* 2012;31:2774–84.
5. Das J, Hammer J, Leonard K. The quality of medical advice in low-income countries. *J Econ Perspect* 2008;22:93–114.
6. Sylvia S, Xue H, Zhou C, *et al*. Tuberculosis detection and the challenges of integrated care in rural China: A cross-sectional standardized patient study. *PLoS Med* 2017;14:e1002405.
7. Hanefeld J, Powell-Jackson T, Balabanova D. Understanding and measuring quality of care: dealing with complexity. *Bull World Health Organ* 2017;95:368–74.
8. Donabedian A. Evaluating the quality of medical care. 1966. *Milbank Q* 2005;83:691–729.
9. Peabody JW, Luck J, Glassman P, *et al*. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA* 2000;283:1715–22.
10. Shah R, Edgar D, Evans BJ. Measuring clinical practice. *Ophthalmic Physiol Opt* 2007;27:113–25.
11. Glassman PA, Luck J, O'Gara EM, *et al*. Using standardized patients to measure quality: evidence from the literature and a prospective study. *Jt Comm J Qual Improv* 2000;26:644–53.
12. Collins J, Harden R. *The use of real patients, simulated patients and simulators in clinical examinations association for medical education in Europe (AMEE) Guide*, 2004.
13. Triola M, Feldman H, Kalet AL, *et al*. A randomized trial of teaching clinical skills using virtual and live standardized patients. *J Gen Intern Med* 2006;21:424–9.
14. Peabody JW, Luck J, Glassman P, *et al*. Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. *Ann Intern Med* 2004;141:771–80.
15. Shah R, Edgar DF, Evans BJ. A comparison of standardised patients, record abstraction and clinical vignettes for the purpose of measuring clinical practice. *Ophthalmic Physiol Opt* 2010;30:209–24.
16. Dresselhaus TR, Peabody JW, Luck J, *et al*. An evaluation of vignettes for predicting variation in the quality of preventive care. *J Gen Intern Med* 2004;19:1013–8.
17. Ellaway R, Candler C, Greene P, *et al*. *An architectural model for MedBiquitous virtual patients*. Baltimore: MedBiquitous, 2006.
18. Cook DA, Triola MM. Virtual patients: a critical literature review and proposed next steps. *Med Educ* 2009;43:303–11.
19. Zhou M, Wang H, Zhu J, *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;387:251–72.
20. Babiarz KS, Miller G, Yi H, *et al*. China's new cooperative medical scheme improved finances of township health centers but not the number of patients served. *Health Aff* 2012;31:1065–74.
21. Li X, Lu J, Hu S, *et al*. The primary health-care system in China. *The Lancet* 2017;390:2584–94.
22. Qian D, Pong RW, Yin A, *et al*. Determinants of health care demand in poor, rural China: the case of Gansu Province. *Health Policy Plan* 2009;24:324–34.
23. Yip WC, Wang H, Liu Y. Determinants of patient choice of medical provider: a case study in rural China. *Health Policy Plan* 1998;13:311–22.
24. Center for Health Statistics and Information. *An analysis report of national health services survey in China*. Beijing: Center for Health Statistics and Information, NHFPC, 2013.
25. Brouwers MC, Kho ME, Browman GP, *et al*. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ* 2010;182:E839–42.
26. Whiting PF, Rutjes AW, Westwood ME, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529–36.

27. Hsu CC, Sandford BA. The delphi technique: making sense of consensus. *Practical Assessment Research & Evaluation* 2007;26:289–304.
28. Zamanzadeh V, Ghahramanian A, Rassouli M, *et al*. Design and implementation content validity study: development of an instrument for measuring patient-centered communication. *J Caring Sci* 2015;4:165–78.
29. Lacasse Y, Godbout C, Sériès F. Health-related quality of life in obstructive sleep apnoea. *Eur Respir J* 2002;19:499–503.
30. Lin LI, Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255–68.
31. McBride G. *A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. NIWA Client Report: HAM2005-062*. Hamilton: National Institute of Water & Atmospheric Research, Ltd (NZ), 2005.
32. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm* 2008;65:2276–84.
33. Rethans JJ, Gorter S, Bokken L, *et al*. Unannounced standardised patients in real practice: a systematic literature review. *Med Educ* 2007;41:537–49.
34. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* 1960;20:37–46.
35. Kwiecien R, Kopp-Schneider A, Blettner M. Concordance analysis: part 16 of a series on evaluation of scientific publications. *Dtsch Arztebl Int* 2011;108:515.
36. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
37. McCabe D, Langer KG, Borod JC, *et al*. Practice effects. In: Kreutzer JS, DeLuca J, Caplan B, eds. *Encyclopedia of clinical neuropsychology*. New York: Springer, 2011:1988–9.
38. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297–334.
39. DeVellis RF. *Scale development: theory and applications*. Los Angeles: Sage, 2012:109–10.
40. Rhodes KV, Miller FG. Simulated patient studies: an ethical analysis. *Milbank Q* 2012;90:706–24.
41. Pharmacotherapy. *Reliability of a virtual patient simulation as an assessment tool*, 2017. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.
42. Urresti-Gundlach M, Tolks D, Kiessling C, *et al*. Do virtual patients prepare medical students for the real world? Development and application of a framework to compare a virtual patient collection with population data. *BMC Med Educ* 2017;17:174.
43. Blok AC, May CN, Sadasivam RS, *et al*. Virtual patient technology: engaging primary care in quality improvement innovations. *JMIR Med Educ* 2017;3:e3.
44. Mollica R, Lavelle J, Fors U, *et al*. Using the virtual patient to improve the primary care of traumatized refugees. *Journal of Medical Education* 2017;16.
45. Mohanan M, Vera-Hernández M, Das V, *et al*. The know-do gap in quality of health care for childhood diarrhea and pneumonia in rural India. *JAMA Pediatr* 2015;169:349–57.