BMJ Open

Assessing the quality of primary healthcare in seven Chinese provinces with unannounced standardised patients: protocol of a cross-sectional survey

Dong Roman Xu,[1] Mengyao Hu,[2] Wenjun He,[3] Jing Liao,[1,3] Yiyuan Cai,[1,3] Sean Sylvia,[4] Kara Hanson,[5] Yaolong Chen,[6] Jay Pan,[7] Zhongliang Zhou,[8] Nan Zhang,[9] Chengxiang Tang,[10] Xiaohui Wang,[11] Scott Rozelle,[12] Hua He,[13] Hong Wang,[14] Gary Chan,[15] Edmundo Roberto Melipillán,[2] Wei Zhou,[16] Wenjie Gong[17]

To cite: Xu DR, Hu M, He W, *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;**9**:e023997. doi:10.1136/bmjopen-2018-023997

► 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-2018-023997).

Received 22 May 2018
Revised 1 November 2018
Accepted 18 December 2018

For numbered affiliations see end of article.

Correspondence to
Professor Wenjie Gong;
gongwenjie@csu.edu.cn

ABSTRACT

Introduction Primary healthcare (PHC) serves as the cornerstone for the attainment of universal health coverage (UHC). Efforts to promote UHC should focus on the expansion of access and on healthcare quality. However, robust quality evidence has remained scarce in China. Common quality assessment methods such as chart abstraction, patient rating and clinical vignette use indirect information that may not represent real practice. This study will send standardised patients (SP or healthy person trained to consistently simulate the medical history, physical symptoms and emotional characteristics of a real patient) unannounced to PHC providers to collect quality information and represent real practice.

Methods and analysis 1981 SP–clinician visits will be made to a random sample of PHC providers across seven provinces in China. SP cases will be developed for 10 tracer conditions in PHC. Each case will include a standard script for the SP to use and a quality checklist that the SP will complete after the clinical visit to indicate diagnostic and treatment activities performed by the clinician. Patient-centredness will be assessed according to the Patient Perception of Patient-Centeredness Rating Scale by the SP. SP cases and the checklist will be developed through a standard protocol and assessed for content, face and criterion validity, and test–retest and inter-rater reliability before its full use. Various descriptive analyses will be performed for the survey results, such as a tabulation of quality scores across geographies and provider types.

Ethics and dissemination This study has been reviewed and approved by the Institutional Review Board of the School of Public Health of Sun Yat-sen University (#SYSU 2017-011). Results will be actively disseminated through print and social media, and SP tools will be made available for other researchers.

Strengths and limitations of this study

- ► We will assess the quality of care with a random sample of primary healthcare providers in seven provinces in China.
- ► We will use unannounced standardised patients (USPs), the ‘gold standard’ of quality assessment.
- ► Both technical quality and patient-centredness will be assessed.
- ► USPs are not suitable for certain health conditions.
- ► The seven provinces are not randomly selected, although we intend for them to represent different health development conditions (using life expectancy as the proxy) in China’s provinces.

BACKGROUND

In 2015, all 191 member states of the United Nations adopted the sustainable development goals, aiming to achieve universal health coverage (UHC)—access to high-quality healthcare services without incurring financial hardship—by 2030.[1] As previous literature emphasised, efforts to promote UHC should focus on the expansion of access and on healthcare quality.[2] Healthcare quality is variously defined by the WHO as the ‘responsiveness’ of the healthcare system to meet desired health outcomes,[3] as the instrumental goals on structure, process and outcome in the Donabedian framework,[4] and as the six comprehensive aims (effectiveness, efficiency, equity, patient-centredness, safety and timeliness) put forth by the Institute of Medicine (IOM).[5] In this study, we adopt the IOM definition of quality.

Primary healthcare (PHC) serves as the cornerstone for the attainment of UHC.[6] China’s latest round of healthcare reform since 2009 has invested heavily in strengthening PHC. There have been some efforts to assess the quality of PHC in China: patients were interviewed with a Primary Care Assessment Tool questionnaire in Guangdong, Shanghai and Hong Kong[7–9]; comprehensiveness of the service provision was used as a proxy for quality through clinician interviewing[10]; and PHC clinicians' adherence to clinical guidelines was assessed with a self-report questionnaire.[11] However, assessment of the quality of PHC has largely remained scant in China, and the assessment tools are indirect and prone to bias.[12] A number of studies have found the quality of PHC to be low in other low-income and middle-income countries (LMICs),[6 13–18] where robust evidence remains scarce.[19] Commonly used methods of measuring technical quality of care include chart abstraction, patient rating of care and using a clinical vignette to test clinician knowledge. Those methods use indirect information that may not represent real practice. This study instead will use unannounced standardised patients (USPs) to measure the quality of real practice. The standardised patient (SP) is a healthy person (or occasionally a real patient) trained to consistently simulate the medical history, physical symptoms and emotional characteristics of a real patient. The SP, particularly when their visit is unannounced, has several reported advantages: (1) reliability in measurement and cross-provider comparison because the same patient is presented to all providers, (2) elimination of the Hawthorne effect (ie, that the study itself may change doctors' behaviour) due to the nature of disguised and unannounced visit by SPs,[20–22] and (3) reduced recall bias.[23 24]

Despite these advantages, the application of SP in China has been concentrated mainly in the area of medical education.[25] An ongoing systematic review identified only four papers on using the SP for quality assessment in China[14 26–28] and 44 in other LMICs. Those projects, often based on a small convenience sample, tended to target a limited number of conditions (approximately 70% on family planning services, childhood infectious diseases, sexually transmitted infections and respiratory tract infections). In this study, we intend to assess the quality of PHC with a probability sample of PHC visits in seven Chinese provinces, using USPs for 10 commonly seen conditions in the PHC setting. The project has involved 20 universities across 19 provinces in China, as well as researchers from Nepal, USA and UK in a USP network (https://www.researchgate.net/project/Unannounced-Standardized-Patient-USP-and-Virtual-Patient-VP-to-Measure-Quality-of-Primary-Care). The USP resources will be pooled and shared widely within the network first and then with the general public. This study is the first of a series of studies to be based on quality data collected using USPs. The primary purpose of this study is to collect and present descriptive data on the quality of China's PHC. We are developing separate protocols for the various hypothesis-driven studies, which will be available elsewhere and from our network website.[29]

METHODS

Survey design

The purpose of the sample design is to create a representative sample of China's PHC providers so that healthcare quality can be assessed based on USP visits to those providers.

Survey population/frame

We considered creating a nationally representative probability sample, but at this stage we have selected seven provinces to 'represent' China due to feasibility considerations. These provinces represent five levels of average life expectancies across China's provinces (figure 1), which are similar to those of five countries with low-income to high-income levels.[30] We intend to create a probability sample that represents PHC in these seven provinces. For the survey population, we intend to include (1) licensed physicians and licensed assistant physicians at community/township health centres/stations and urban health stations; (2) certified village doctors (a terminology in China that refers to village clinicians who have village-level practice privilege even without a medical licence) and village sanitarians (referring to uncertified village doctors who are supposed to work under the supervision of the village doctor) at village clinics; and (3) clinicians with a licence notation for general practice, internal medicine, obstetrics/gynaecology and paediatrics at the level I and level II hospitals and the maternal and childcare centres. We exclude level 3 hospitals, which provide more specialised care, and specialty hospitals. Clinicians meeting those criteria will constitute the 'sampling frame'.

Sampling procedures

The sample will be selected using a multistage, clustered sample design covering all eligible clinicians in the seven provinces (figure 2). In the first stage, stratification will be based on the provinces. Due to the high number of visits in the seven capital cities, we will sample each capital city. Each province is thus divided into two strata consisting of the provincial capital city and other prefecture-level municipalities, leading to 14 strata in total. We will use proportionate allocation (in terms of the number of eligible clinicians) of the sample size for each stratum. For each stratum, five rural townships or urban subdistricts (the primary sampling unit [PSU]) will be selected using probability proportional to size (PPS). In the second stage, for each PSU, PHC facilities as previously defined (secondary sampling unit [SSU]) will be selected using PPS systematic sampling. Neighbouring village clinics will be grouped as an SSU. The number of SSUs for each stratum will vary depending on the size of the stratum—for example, more SSUs will be selected in strata with more PHC clinicians. In the final stage, a fixed number of USP visits will be made to each selected facility or the group of facilities in the case of village clinics. The exact number of visits will be determined once we obtain and examine our sampling frame.

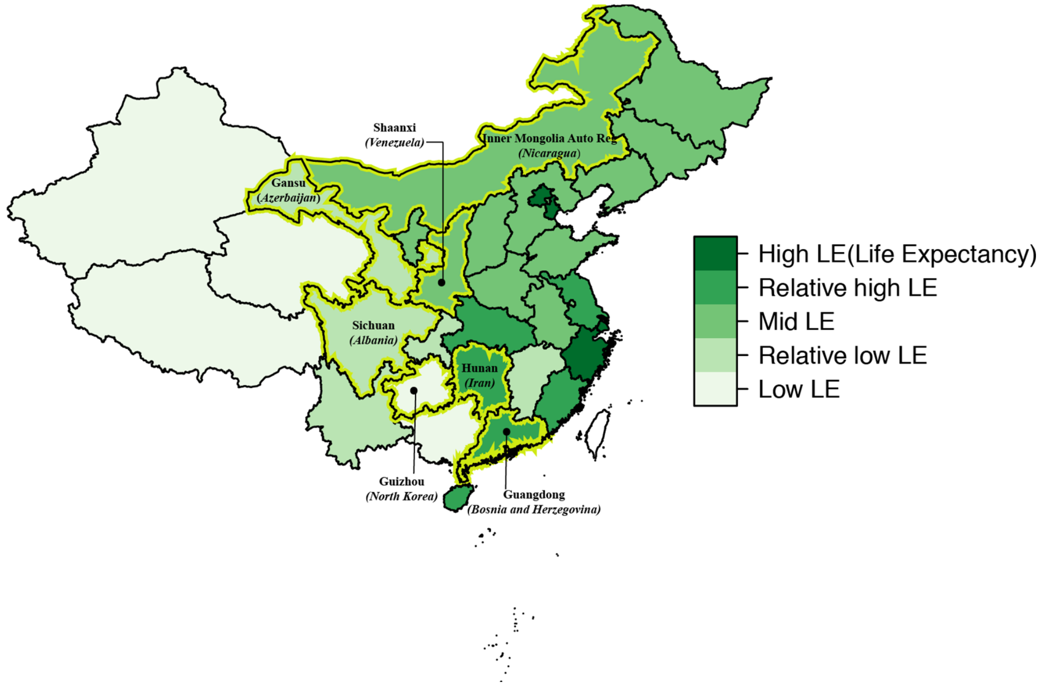

Figure 1 Seven selected sample provinces on the map of China with referencing countries of equivalent life expectancy in brackets. The figure is adapted from the paper by Liao *et al*[29]. Permission to use has been obtained.

Sample size calculation

The sample size was calculated for the primary purpose of the standard descriptive survey analysis of this survey. The sample size (power) calculation for other related hypotheses of related studies will be described in separate study protocols. The primary statistic of interest in this survey is a latent variable measuring clinicians' quality, constructed using the two-parameter logistic item response theory (IRT) model.[31 32] The model was based on a list of quality checklist items measuring whether doctors asked recommended questions and whether they performed recommended exams (see the Scoring methods section below). Survey sample size was calculated based on the desired level of relative precision (coefficient of variation, CV), an estimate for the population element variance for the variable of interest (s^2) from previous study and design effect (*deff*). In this study, our desired level of relative precision (CV) is 0.08. s^2 was estimated to be 4.54, based on Sylvia *et al*'s[14 27] work on the USP-assessed quality of PHC in three Chinese provinces. Design effect is the variance inflation due to cluster sampling. This figure was calculated based on intraclass correlation (ICC) (describing the level of homogeneity of the units in a cluster) and cluster sample size: $deff = 1 + \delta\left(n - 1\right)$, where δ is the ICC and n is the average size of the cluster. The ICC of 0.0486 was also estimated from Sylvia *et al*'s work. Our estimated average cluster size is 27 clinician–SP encounters per PSU. Accordingly, we calculated the total required sample size to be 1981 clinician–SP encounters. The steps taken to calculate the sample size can be found in online supplementary appendix 1.

USP case development and implementation

The development process of a USP case is based on our extensive literature review[20 33] as well as our own USP experiences in Shaanxi Province, China.[14 27] We are concurrently developing smartphone-based virtual standardised patients (VPs) (details described elsewhere).

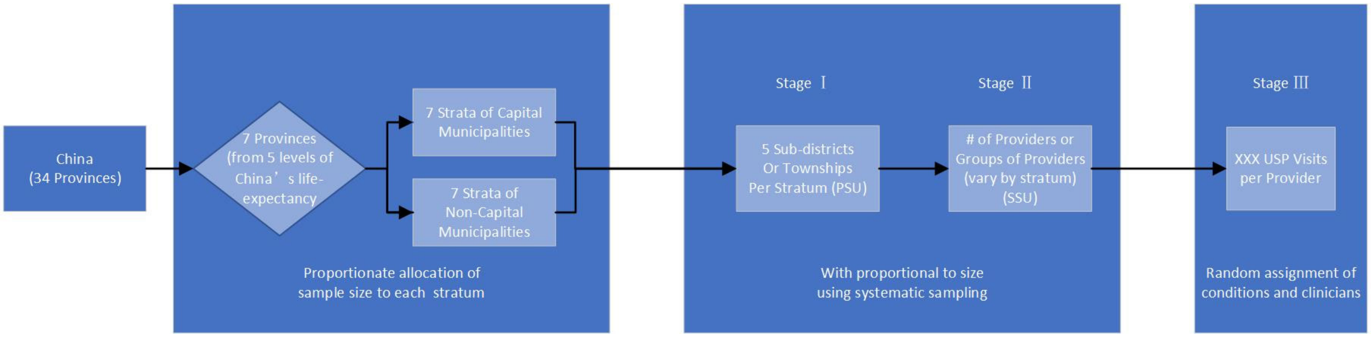

Figure 2 Sampling procedure. PSU, primary sampling unit; SSU, secondary sampling unit; USP, unannounced standardised patient.

The two projects will share almost identical case scenarios and quality criteria.

Case selection

Our purpose is to select 10 health problems as tracer conditions for PHC in China. Ideally our selected cases should (1) be highly prevalent in PHC settings; (2) carry challenging features in different aspects of PHC (eg, some cases focus on curative care, while others on prevention, disease management, culturally sensitive care[34] or misuse of low-value tests[35–37]); (3) not involve invasive and painful procedures; and (4) not require physical signs that cannot be simulated (eg, jaundice can be simulated with make-up, but heart murmurs cannot).[23] We created a list of the top 30 conditions commonly seen in PHC in China, combining the results of two national surveys on PHC.[12] A panel of physicians and public health and health system researchers then applied the principles above and selected a dozen of PHC problems for USP development (table 1). Ten final conditions will be selected from this list.

Development team

We have created an overall development team and 10 case-specific development teams. Each team includes case-specific specialists, general practitioners, and public health and health system researchers (online supplementary appendix 2). A third overall panel consisting of primary care providers at the village, township and community levels will review all cases for contextual appropriateness in primary care settings. In developing the case, we will follow several principles: (1) limiting case scenarios to those that require definitive clinician action on the first visit to minimise potential 'first-visit bias',[38] (2) focusing on the presentation of symptoms for which evidence is well established for diagnosis and management, and (3) deriving some content of the cases from the actual case history of relevant patient files in real practice.[23]

Case description

The case description describes the relevant clinical roles and psychosocial biographies of the SP.[39] We used a structured description of the cases as follows:

1. Social and demographic profile: (1) socioeconomic information: name, gender, age, ethnicity, education, occupation, family structure (eg, married and have two children but live alone), dress style (eg, dressed in jeans, work boots and a well-worn but neat sweater), health insurance or other social programme participation; (2) personality that may influence interaction with the clinician (eg, non-proactive and introverted); and (3) lifestyle relevant to health (eg, smoked one pack of cigarette since age 18, like fried pork but also eat much fruit, exercise regularly, watch television a lot during spare time, play mah-jong with friends and visit children every week).
2. Medical history: (1) disease information: severity of the condition (eg, mild or severe depression), duration of the condition (the first onset? previously diagnosed/existing [how long?]), comorbidity (any other physical and/or psychological problems?); (2) reason for seeking care for this specific visit (eg, was feeling down for 2 months but depression worsened last week); and (3) treatment/management already or currently received (eg, a 'patient' with diabetes took metoprolol for hypertension but does not monitor his glucose/watch his diet/weight).
3. Physical examination: symptoms the SP will (and will not) portray (eg, reduced appetite, but not showing agitation), and medical signs the SP has or does not have (eg, heart murmur).
4. Laboratory and imaging: laboratory and imaging that a clinician may prescribe for the SP. The laboratory and imaging results of the SP may be generated from those of real typical patients.
5. Diagnosis: the correct diagnosis that the clinician should make based on the information presented by the SP.
6. Treatment and management: the decision of the clinician on what medications, procedures, advice or referral will be given at the end of the consultation.

Script

Corresponding to the six components of the aforementioned case description, we will develop a detailed script for the SPs to use in their PHC visit with the clinician. The script ideally should cover all possible questions a clinician may ask, as well as the SP's answers during the clinical interaction. Panels of clinicians will be consulted to collect relevant questions that will guide the development of the script. The script will continue to add new questions asked by the clinicians on the SP–clinician interaction. The script will have five sections: (1) an opening: spontaneous information given to the clinician at the start (eg, Doctor, I have had a headache for 2 days), (2) the information given only on request, (3) the information for the SP to volunteer even if not asked, (4) the language to insist on a diagnosis if not given and (5) an ending.[14 20 40]

Quality checklist

The checklist consists of explicit quality criteria for gathering data on patient history, physical examination, laboratory/imaging, diagnosis and treatment.[14 33] Based on our comprehensive review of 14 articles on literature and evidence-based clinical guideline development methodology,[41] we have established a guiding principle and standard protocol for checklist development. Our process will (1) be evidence-based and augmented by expert opinion,[42] (2) follow a systematic procedure to gather, evaluate and select evidence and criteria, (3) select criteria related to clinician actions that the SP can easily evaluate,[43] and (4) keep the number of checklist items under 30 to include high-priority criteria only so that the

Table 1 Selected candidate conditions

		Special focus areas											
	Conditions	Chronic disease management	Public health delivery	Mental health	Maternal and childcare	Preventative care	Referral	Patient-centred care	Older adults	Low-value diagnostic	Antibiotics	Process traditional Chinese drug	Injury
1	Common cold (influenza season)					X					X	X	
2	Hypertension	X										X	
3	Type 2 diabetes mellitus	X						X	X			X	
4	Gastritis							X					
5	Child diarrhoea				X						X		
6	Low back pain (patient requesting low-value test)							X		X			
7	Depression (maternal care)			X	X		X	X					
8	Angina (heavy smoker)					X	X	X				X	
9	Headache											X	
10	Fall					X		X	X				X
11	Asthma												
12	Tuberculosis		X			X	X						

SP can reliably recall clinician behaviour.[43–45] The details of our checklist development protocol will be described in a separate paper, and key messages are summarised in online supplementary appendix 2.

Selecting and training SPs

We will advertise on social media to recruit SPs. The candidate must be in stable health without confounding symptoms; should match the real patients in age, sex and physical features; are willing to allow the examinations appropriate to their condition; and have the intellectual maturity to present the behaviour of the actual patient and complete the checklist.[23 46 47] We may consider recruiting real patients with stable conditions to portray the cases not subject to simulation.[23] The training of the SP will aim at portraying the signs, symptoms and presentations, completing the checklist, and minimising detection by the provider.[20] The week-long training will have three stages: classroom instruction, a dress rehearsal and two field tests.[23 47 48] Each case will have three SPs who will be trained according to a standardised training manual that will be developed to guide the training and appraisal of the SPs.

Fielding and implementing SPs

A disguise plan will be developed for each case to minimise physician detection of the SP status (eg, convincing excuse for seeking care where they do not usually reside). In the pilot (instrument validation) phase, consent will be sought for audio recording (see below); in these cases, fieldwork will start only 3–4weeks after consent is obtained. We will provide each SP with a calamity letter, explaining the project in case of their identity being exposed.

After the facilities are selected, and the number of visits per facility is determined, each of the planned visits will be given a unique identifier (eg, facility A-1, facility A-2, facility B-1), which will then be randomly ordered to form a random sequence numbered from 1 to 1981 consecutively. One of the ten SP cases will be randomly assigned to each number on this random sequence. The seven SPs per case will be dispatched to the seven provinces concurrently, one SP per province. If multiple clinicians are available in that facility at the time of a particular SP visit (PHC visits in China do not require appointments), the field coordinator will randomly select a clinician by drawing lots onsite. Each SP is expected to make a total of approximately 30 visits. We plan to complete those SP visits over a 3-month time span.

In a separate but related study, a week after the visit of the SP, the same clinician will perform the same consultation but with a standardised virtual patient on a smartphone.[29] We will use this opportunity to administer a detection questionnaire to the clinician, asking whether they suspect they had any visit from an SP over the past week. The detected cases will be treated as missing data in the data analysis.

Variables

Outcome variables

We will collect a variety of quality of care information and other related explanatory variables. The IOM quality framework (effective, safe, patient-centred, timely, efficient and equitable) will be used for quality evaluation (table 2). *Effectiveness* (avoiding underuse and misuse) and *safety* (avoiding harm), traditional technical goals of quality of care, will be evaluated through the yes/no checklist discussed above (online supplementary appendix 2). *Patient-centredness* (respectful of and responsive to individual preferences) will be assessed by the Patient Perception of Patient-Centeredness (PPPC) Rating Scale.[49–51] Using a 4-point Likert scale, the PPPC Rating Scale evaluates three dimensions of patient-centredness: exploring the disease and illness experiences, understanding the whole person and finding common ground.[49] Prior studies have demonstrated the validity of SPs rating clinician communications.[52 53] A separate study will be conducted to test the validity of the PPPC Rating Scale. *Timeliness* will be assessed by analysing opening hours, waiting time and consultation time.[5] *Efficiency* (avoiding waste) will be measured by costs of care of the SP–clinician encounter. *Equity of care* (no variance in quality because of personal characteristics) will be assessed through a separate but related study in a randomised cross-over trial.

Scoring methods

Technical quality of care will be reflected by a continuous score ranging from 0 to 1. We will evaluate further whether to classify checklist items in four categories (essential, important, indicated and non-contributory) with corresponding numeric weights (3, 2, 1 and 0).[54] Two scoring methods will be used: (1) the simple scoring method will use the formula of items performed divided by the total number of items on the checklist for the process scores, whereas (2) the complex method will use an algorithm based on the IRT.[31] Using the IRT model approach, we can obtain a latent performance score for each doctor, which has been corrected for measurement error. An ordinal variable will be used for diagnosis and management plans (table 2), while patient-centredness will follow the scoring methods of the PPPC Rating Scale (possible range of score from 1 to 4).[51]

Other variables

We will collect additional information on the predictors, confounders and effect modifiers to the outcomes in the planned hypothesis testing of the related studies to this survey. The information will include qualifications of the clinician and facility information (environment, amenity, size, location, ownership type and so forth).

Analytical methods

USP validation

USP validation will be based on a convenience sample of clinicians not included in our final survey sample in

Table 2 Variables

	Variable name	Type	Coding	Source
1. Effectiveness and safety				
1.1	% of recommended questions asked	Continuous	0–1	SP checklist
1.2	% of recommended exams performed	Continuous	0–1	SP checklist
1.3	Diagnosis quality	Ordinal	0: incorrect, 1: partially correct, 2: correct	SP checklist
1.4	Treatment quality	Ordinal	0: incorrect, 1: partially correct, 2: correct	SP checklist
2. Patient-centredness				
2.1	Patient perception of patient-centredness	Continuous	0–1	PPPC
2.2	Choice of provider	Dichotomous	0: no, 1: yes	SP checklist
2.3	Ease of navigation in facility	Ordinal	0: difficult, 1: median, 2: easy	SP rating
3. Timeliness				
3.1	Opening hours	Continuous	Hours	SP checklist
3.2	Wait time	Continuous	Minutes	SP checklist
3.3	Consultation time	Continuous	Minutes	SP checklist
4. Efficiency				
4.1	Total cost	Continuous	Renminbi	SP checklist
4.2	Medication cost	Continuous	Renminbi	SP checklist
4.3	Laboratory/imaging cost	Continuous	Renminbi	SP checklist
5. Equity				
5.1	To be analysed in a separate cross-over trial			

PPPC, Patient Perception of Patient-Centeredness Rating Scale; SP, standardised patient.

the project training and pilot phase. Those SP–clinician interactions in the pilot will be audio-recorded and transcribed. *Validity* is the extent to which an instrument measures what it is supposed to measure. We will assess content, face and criterion validity of the cases. Content validity will be assessed by an expert panel who will use a 4-point Likert scale to evaluate the appropriateness of the written content of the cases that will include the scenario, scripts and checklists. For the checklist, they will be instructed to check the appropriateness against the published clinical guidelines. The face validity of the SP assessment depends on (1) the SP remaining undetected (detection ratio reported to be 5%–10%[55]), and (2) authentically and consistently portraying the clinical features of the case. We will send the participating clinician in the pilot a 'detection form' to report their degrees of suspicion of any SP visit.[46] The authenticity of the SP presentation will be evaluated by checking the transcribed recording to discover whether a key piece of information was divulged by the SP when appropriately prompted, not divulged when prompted or volunteered when not prompted. Criterion validity will be assessed through the agreement of the SP-completed checklist against that completed by a clinician based on the transcript of the visit (ie, the clinician rating as the 'gold standard').[56–59] Checklist items which depend on visual observation will be excluded. *Reliability* examines the level of consistency of the repeated measurements. The inter-rater reliability of two SPs on the same condition and context will be assessed with two SPs completing the checklist for the same recorded transcript. Test–retest reliability will be analysed by the concordance of assessment results of the same SP to score his or her own recorded encounter a month later.[57] The agreement will be analysed with Lin's concordance correlation coefficient (r_c).[60] r_c indicates how closely pairs of observation fell on a 45° line (the perfect concordance line) through the origin in addition to their correlation.[60–62] Bland-Altman plot will be used to visualise the concordance.[63 64] Table 3 summarises our methods of validation.

Survey analysis

We will focus on descriptive analysis to present the quality of PHC in the seven provinces. Hypothesis-driven analyses will be described in separate study protocols. For descriptive analysis, we will first present clinician and facility profiles in tables for all seven provinces and by each province. The clinician profile will include sociodemographic information (age, gender and ethnicity), professional qualification (general and medical education, licensure, and professional ranks) and service information (volume of visits and number of support personnel). The

Table 3 Methods of validation for the USP cases

Domain	Indicator	Data collection		Statistical analysis
		Phase	Method	
Content validity	Content Validity Index (CVI)	USP case review	Expert panel review of SP cases, measured by a 4-point Likert scale (1=lowest, 4=highest).	CVI for SP case and for specific USP, where CVI=number of raters giving a rating of 3 or 4 divided by the total number of raters.
Face validity	Authenticity of SP role-play	Validation study	Transcripts of the recording of the USP–clinician encounter to be assessed by a member of the project team for accuracy of portraying the clinical case by a 5-point Likert scale (1=100% inaccurate, 5=100% accurate).	Accuracy score=per cent of positive evaluations (ie, evaluation ≥4).
	Detection ratio		Clinicians receiving an SP visit to complete a 'detection form' afterwards to report any suspected USP visits: 0=not suspected; 1=somehow suspected; 2=suspected with certainty).	Detection ratio=number of detected USP visit divided by the total number of USP visits (for case-specific detection ratio and all-case detection ratio, respectively). Detection ratio of 10% and less is considered acceptable.
Criterion validity	Lin's concordance correlation coefficient (r_c); kappa statistic	Validation study	SP-completed checklist against that by a clinician based on the transcript of the visit (ie, the clinician rating as the 'reference standard').	The concordance of the quality scores based on SP-completed checklist against that based on the reference standard. r_c used for continuous process quality scores, and kappa for dichotomous diagnoses and treatment and management measures.
Test–retest reliability	Lin's concordance correlation coefficient (r_c); kappa statistic	Validation study.	The same SP to score his own recorded encounter in a month.	The concordance to 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.
Inter-rater reliability			Multiple SPs to complete the checklist for the same recorded transcript.	

SP, standardised patient; USP, unannounced standardised patient.

facility profile will include information on operation and management (years in operation, ownership types, accreditation, level of hospitals, affiliation with medical universities, revenue, health insurance contracting, payment methods), clinical services (annual number of inpatient and outpatient visits, number of clinical departments), personnel (number of physicians, nurses and attrition ratio) and equipment. Second, we will tabulate the results of overall quality and subdomains across administrative regions and provider types. Third, we will map out the locations of the facilities along with their quality scores with geospatial analytical tools. Finally, a t-test/Wilcoxon test or χ^2 test will be employed to compare quality differences between public versus private providers, primary care clinics/centres versus hospital outpatient services, care in rural versus urban areas, and across different conditions, clinician educational levels and payment mechanisms.

Related studies

This study protocol mainly deals with the descriptive analysis and presentation of the data to be collected by the

USPs. Using the USP survey data, we have planned several related studies that will be covered by separate study protocols with details on the background, theoretical framework and analytical methods. To summarise those related studies, we will assess (1) the effect of ownership types of the PHC providers (ie, private vs public) on the quality of PHC (study protocol under revision), (2) the know-do gap between the assessment results by a smartphone-based VPs and USP (protocol already published),[29] (3) the effect of using smartphone-based virtual patient in improving clinician performance, (4) the effect of types of insurance carried by a patient on quality of care, (5) the impact of gatekeeping by primary care providers on quality of tuberculosis care—a mathematical modelling study, and (6) clinician skills in handling low-value or harmful patient-requested services, particularly antibiotics and some processed traditional Chinese medicine.

Ethics and dissemination

USP studies do not necessarily require consent if they meet certain conditions.[65 66] Our waiver has been granted for the following reasons: (1) our study serves important public good, while requiring informed consent may lead to considerable selection bias and greater risk for the detection of the SP; (2) this study does not intend to entrap or reveal identities of any institution or individual, and all analyses will be conducted at the broader health system level (after data cleaning, all individual identifiers will be destroyed); and (3) no audiovisuals will be recorded during the SP–clinician encounter (however, in the pilot stage, we will seek informed consent from participating clinicians as we will use a disguised recording for the validation purposes). The study results will be widely distributed in the form of scientific papers and policy briefs. The data generated from this project and the USP cases and accompanying user manuals will be made available to other researchers on request after we complete our primary analysis.

Patient and public involvement

We selected the conditions for the USP partly based on results from surveys on common conditions in the context of PHC as reported by patients. The USP cases will also be reviewed by a panel that includes patients. The results of the studies will be widely distributed in scientific reports as well as social media to benefit policymakers, clinicians and patients.

DISCUSSION

In this study, we will develop, validate and implement methods of assessing the quality of PHC using USPs. Compared with existing studies using USPs,[33] this proposed study has several distinctive features. First, we will establish a large probability random sample so that representative estimates of PHC quality can be achieved in the chosen seven provinces in China. Second, unlike previous studies,[14 27] we include village clinics, township health centres and community health centres, and also county hospitals and other level I and level II hospitals, in the study. The latter were not officially designated as PHC facilities in China but provided a substantial amount of PHCs. Third, 10 SP cases will be developed through a standardised process using the same template and methodology and will represent common conditions in PHC, while past studies often used two to three conditions.[33] Fourth, an evidence-based systematic method will guide checklist development. In a review, only 12 out of 29 SP articles reported the procedures of checklist development and many checklists were developed by expert consensus only.[54] Fifth, in addition to using the checklist to evaluate technical quality of care as performed in most other USP studies, we will assess patient-centredness with a global rating scale. Sixth, we have planned a series of related studies to address the quality of PHC in a concerted effort. Most noteworthy, we are developing 10 identical conditions as smartphone-based virtual patients to assess the competency of PHC providers. Seventh, we used the same case for all levels of providers, from village doctors to township health centres, to county hospitals, but quality checklists for process, diagnosis and treatment will be tailored to fit the expected roles and responsibilities of the different providers. Finally, we have secured the understanding and cooperation of the provincial health authorities.

We note two particular issues. In high-income settings, logistical arrangements for the SP are complex. A significant challenge is to introduce the SP into medical practice.[23 47 48] However, in China and many other LMICs, enrolment with a clinician is not required, and a walk-in visit to clinicians without an appointment is commonplace. However, village doctors usually know their patients well. For these areas, the SPs in other studies pretended to be tourists or friends visiting the families in the village. We will try other pretences, such as a temporary poverty-relief worker who has just arrived in a nearby village. Those poverty-relief workers are common in remote rural areas in China. For the second issue, assessing quality with USP was reported to incur high cost in developed countries (estimated to be US$350–400 per visit).[53 67] We expect the cost in China to be considerably lower due to the lower labour cost. We will collect detailed cost information to inform the future application of the USP.

The study has several potential limitations. Most important, even though the assessment of SP is considered the gold standard for measuring clinician performance, and in this study we have further expanded the use of SPs to evaluate other elements of quality in the IOM framework such as patient-centredness, timeliness and efficiency, we recognise that those quality of care elements are still largely clinician-related, and other important quality aspects such as the quality of laboratory testing cannot be assessed by our SPs. In addition, the USP method has several technical challenges. If healthy people are used to simulate the patient, it is difficult to achieve complete alignment of patient presentation of signs and symptoms (for instance, it is difficult to fake a sore throat). There are also challenges to obtaining fake laboratory test results that may be necessary for the diagnosis. Some clinical roles that require the SP to go through invasive investigation may also pose a problem. We will experiment with a real patient in stable conditions to resolve some of those challenges. Next, our judgement of the clinical quality through the first and only visit with the SP may lead to 'first-visit bias'.[38] The quality of care provided by a clinician who spreads his or her diagnosis and management over several visits may be underestimated. We try to minimise this bias by designing cases that require a definitive decision on the first visit. Last, even though we intend to select 10 tracer conditions in the context of PHC, we still need to be cautious in generalising the findings to the overall quality of PHC.

In conclusion, this proposed study may produce a set of validated tools for the assessment of the quality of PHC using USP and apply it to obtain valuable quality of care information on PHC in China.

Author affiliations

[1]Sun Yat-sen Global Health Institute (SGHI), School of Public Health and Institute of State Governance, Sun Yat-sen University, Guangzhou, China
[2]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA
[3]Department of Biostatistics and Epidemiology, School of Public Health, Sun Yat-sen University, Guangzhou, China
[4]Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[5]Department of Global Health and Development, Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UK
[6]Evidence Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou, China
[7]West China School of Public Health, Sichuan University, Chengdu, Sichuan, China
[8]School of Public Policy and Administration, Xi'an Jiaotong University, Xi'an, China
[9]Department of Health Management, School of Health Management, Inner Mongolia Medical University, Hohhot, China
[10]School of Public Administration, Guangzhou University, Guangzhou, China
[11]Department of Social Medicine and Health Management, School of Public Health, Lanzhou University, Lanzhou, Gansu, China
[12]Freeman Spogli Institute for International Studies, Stanford University, Stanford, California, USA
[13]Department of Epidemiology, School of Public Health and Tropical Medicine, Tulane University, New Orleans, USA
[14]Health Economics, Financing and Systems, Bill & Melinda Gates Foundation, Seattle, USA
[15]Department of Biostatistics, University of Washington, Seattle, Washington, USA
[16]Hospital Administration Institute, Xiangya Hospital, Central South University, Changsha, China
[17]Xiangya School of Public Health, Central South University, Changsha, China

Contributors DRX conceived the project concept and developed the first protocol draft, along with WG. DRX, MH and WH developed the sampling design, and MH, WH and ERM wrote the section on samples and performed the sample size calculation. SS provided original data of the previous studies for the sample size estimation and calculated some summary statistics. JL and Y-LC worked on the SP case templates. Y-LC and XW developed the guideline for the development of the quality checklist. KH reviewed the content and edited the manuscript. HH and GC reviewed the statistical plan. SR, JP, HW, ZZ, CT, NZ and WZ reviewed and commented on the design and methods. All coauthors participated in the revision and approved this manuscript.

Funding This project is funded through the following competitive grants: China Medical Board (grant number: CMB GNL 16-260), National Natural Science Foundation of China (grant number: 81773446) and Young Scientists Fund of the National Natural Science Foundation of China (grant number: 81402690).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This study has received ethical approval from the institutional review board (IRB) of the Sun Yat-sen University School of Public Health with a waiver of informed consent from each participating clinician.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. A/RES/70/1 R. Transforming our world: the 2030 agenda for sustainable development. 2018. http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E (accessed 17 Feb 2018).
2. Hanefeld J, Powell-Jackson T, Balabanova D. Understanding and measuring quality of care: dealing with complexity. *Bull World Health Organ* 2017;95:368–74.
3. Murray CJ, Frenk J. *A WHO framework for health system performance assessment: Evidence and Information for Policy*: World Health Organization, 1999.
4. Donabedian A. The quality of care. How can it be assessed? 1988. *Arch Pathol Lab Med* 1997;121:1145.
5. Pongsupap Y, Van Lerberghe W. Choosing between public and private or between hospital and primary care: responsiveness, patient-centredness and prescribing patterns in outpatient consultations in Bangkok. *Trop Med Int Health* 2006;11:81–9.
6. Bitton A, Ratcliffe HL, Veillard JH, *et al*. Primary health care as a foundation for strengthening health systems in low- and middle-income countries. *J Gen Intern Med* 2017;32:566–71.
7. Wei X, Li H, Yang N, *et al*. Changes in the perceived quality of primary care in Shanghai and Shenzhen, China: a difference-in-difference analysis. *Bull World Health Organ* 2015;93:407–16.
8. Zou Y, Zhang X, Hao Y, *et al*. General practitioners versus other physicians in the quality of primary care: a cross-sectional study in Guangdong Province, China. *BMC Fam Pract* 2015;16:134.
9. Feng S, Shi L, Zeng J, *et al*. Comparison of primary care experiences in village clinics with different ownership models in Guangdong Province, China. *PLoS One* 2017;12:e0169241.
10. Wong WCW, Jiang S, Ong JJ, *et al*. Bridging the gaps between patients and primary care in china: a nationwide representative survey. *Ann Fam Med* 2017;15:237–45.
11. Zeng L, Li Y, Zhang L, *et al*. Guideline use behaviours and needs of primary care practitioners in China: a cross-sectional survey. *BMJ Open* 2017;7:e015379.
12. Li X, Lu J, Hu S, *et al*. The primary health-care system in China. *The Lancet* 2017;390:2584–94.
13. Das J, Hammer J. Quality of primary care in low-income countries: facts and economics. *Annu Rev Econom* 2014;6:525–53.
14. 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.
15. Berendes S, Heywood P, Oliver S, *et al*. Quality of private and public ambulatory health care in low and middle income countries: systematic review of comparative studies. *PLoS Med* 2011;8:e1000433.
16. 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.
17. Das J, Gertler PJ. Variations in practice quality in five low-income countries: a conceptual overview. *Health Aff* 2007;26:w296–309.
18. Das J, Hammer J, Leonard K. The quality of medical advice in low-income countries. *J Econ Perspect* 2008;22:93–114.
19. Coarasa J, Das J, Gummerson E, *et al*. A systematic tale of two differing reviews: evaluating the evidence on public and private sector quality of primary care in low and middle income countries. *Global Health* 2017;13:24.
20. 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.
21. Leonard K, Masatu MC. Outpatient process quality evaluation and the Hawthorne Effect. *Soc Sci Med* 2006;63:2330–40.

22. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J Clin Epidemiol* 2014;67:267–77.
23. Woodward CA, McConvey GA, Neufeld V, *et al*. Measurement of physician performance by standardized patients. Refining techniques for undetected entry in physicians' offices. *Med Care* 1985;23:1019–27.
24. Das J, Hammer J. Money for nothing: the dire straits of medical practice in Delhi, India. *J Dev Econ* 2007;83:1–36.
25. Yu-jie Z, Min W, Qin L. Analyze the development of standardized patient teaching in China by literature review in recent 10 years. *Chin J Nurs* 2009;44:259–61.
26. Currie J, Lin W, Zhang W. Patient knowledge and antibiotic abuse: Evidence from an audit study in China. *J Health Econ* 2011;30:933–49.
27. 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.
28. Li L, Lin C, Guan J. Using standardized patients to evaluate hospital-based intervention outcomes. *Int J Epidemiol* 2014;43:897–903.
29. 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.
30. 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. *The Lancet* 2016;387:251–72.
31. Das J, Hammer J. Which doctor? Combining vignettes and item response to measure clinical competence. *J Dev Econ* 2005;78:348–83.
32. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*. Sage, 1991.
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. Kutob RM, Bormanis J, Crago M, *et al*. Assessing culturally competent diabetes care with unannounced standardized patients. *Fam Med* 2013;45:400–8.
35. Fenton JJ, Kravitz RL, Jerant A, *et al*. Promoting patient-centered counseling to reduce use of low-value diagnostic tests: a randomized clinical trial. *JAMA Intern Med* 2016;176:191–7.
36. May L, Franks P, Jerant A, *et al*. Watchful Waiting Strategy May Reduce Low-Value Diagnostic Testing. *J Am Board Fam Med* 2016;29:710–7.
37. Zabar S, Hanley K, Lee H, *et al*. Ordering of labs and tests: variation and correlates of value-based care in an unannounced standardized patient visit. *J Gen Intern Med* 2016;32:S318.
38. Tamblyn RM, Abrahamowicz M, Berkson L, *et al*. First-visit bias in the measurement of clinical competence with standardized patients. *Acad Med* 1992;67:S22–4.
39. Shepherd HL, Barratt A, Trevena LJ, *et al*. Three questions that patients can ask to improve the quality of information physicians give about treatment options: a cross-over trial. *Patient Educ Couns* 2011;84:379–85.
40. Peabody JW, Luck J, Jain S, *et al*. Assessing the accuracy of administrative data in health information systems. *Med Care* 2004;42:1066–72.
41. Organization WH. *WHO handbook for guideline development*: World Health Organization, 2014.
42. Campbell SM, Braspenning J, Hutchinson A, *et al*. Research methods used in developing and applying quality indicators in primary care. *Qual Saf Health Care* 2002;11:358–64.
43. De Champlain AF, Margolis MJ, King A, *et al*. Standardized patients' accuracy in recording examinees' behaviors using checklists. *Acad Med* 1997;72:S85–7.
44. Vu NV, Steward DE, Marcy M. An assessment of the consistency and accuracy of standardized patients' simulations. *J Med Educ* 1987;62:1000–2.
45. Vu NV, Marcy MM, Colliver JA, *et al*. Standardized (simulated) patients' accuracy in recording clinical performance check-list items. *Med Educ* 1992;26:99–104.
46. Maiburg BH, Rethans JJ, van Erk IM, *et al*. Fielding incognito standardised patients as 'known' patients in a controlled trial in general practice. *Med Educ* 2004;38:1229–35.
47. Gorter SL, Rethans JJ, Scherpbier AJ, *et al*. How to introduce incognito standardized patients into outpatient clinics of specialists in rheumatology. *Med Teach* 2001;23:138–44.
48. Siminoff LA, Rogers HL, Waller AC, *et al*. The advantages and challenges of unannounced standardized patient methodology to assess healthcare communication. *Patient Educ Couns* 2011;82:318–24.
49. Oates J, Weston WW, Jordan J. The impact of patient-centered care on outcomes. *Fam Pract* 2000;49:796–804.
50. Hudon C, Fortin M, Haggerty JL, *et al*. Measuring patients' perceptions of patient-centered care: a systematic review of tools for family medicine. *Ann Fam Med* 2011;9:155–64.
51. Brown J, Stewart M, Tessier S. *Assessing communication between patients and doctors: a manual for scoring patient-centred communication*. London: Thames Valley Family Practice Research Unit, 1995.
52. Ozuah PO, Reznik M. Can standardised patients reliably assess communication skills in asthma cases? *Med Educ* 2007;41:1104–5.
53. Zabar S, Ark T, Gillespie C, *et al*. Can unannounced standardized patients assess professionalism and communication skills in the emergency department? *Acad Emerg Med* 2009;16:915–8.
54. Gorter S, Rethans JJ, Scherpbier A, *et al*. Developing case-specific checklists for standardized-patient-based assessments in internal medicine: a review of the literature. *Acad Med* 2000;75:1130–7.
55. Franz CE, Epstein R, Miller KN, *et al*. Caught in the act? Prevalence, predictors, and consequences of physician detection of unannounced standardized patients. *Health Serv Res* 2006;41:2290–302.
56. Swartz MH, Colliver JA, Bardes CL, *et al*. Validating the standardized-patient assessment administered to medical students in the New York City Consortium. *Acad Med* 1997;72:619–26.
57. Rethans JJ, Drop R, Sturmans F, *et al*. A method for introducing standardized (simulated) patients into general practice consultations. *Br J Gen Pract* 1991;41:94–6.
58. Luck J, Peabody JW. Using standardised patients to measure physicians' practice: validation study using audio recordings. *BMJ* 2002;325:679.
59. Shirazi M, Sadeghi M, Emami A, *et al*. Training and validation of standardized patients for unannounced assessment of physicians' management of depression. *Acad Psychiatry* 2011;35:382–7.
60. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255–68.
61. Steichen TJ, Cox NJ. A note on the concordance correlation coefficient. *Stata J* 2002;2:183–9.
62. Lawrence I, Lin K. Assay validation using the concordance correlation coefficient. *Biometrics* 1992:599–604.
63. 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.
64. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
65. Rhodes K. Taking the mystery out of "mystery shopper" studies. *N Engl J Med* 2011;365:484–6.
66. Rhodes KV, Miller FG. Simulated patient studies: an ethical analysis. *Milbank Q* 2012;90:706–24.
67. Weiner SJ, Schwartz A. Directly observed care: can unannounced standardized patients address a gap in performance measurement? *J Gen Intern Med* 2014;29:1183–7.