How to do (or not to do)… using the standardised patient method to measure clinical quality of care in LMIC health facilities

Abstract

Standardised patients, that is, mystery shoppers for healthcare providers, are increasingly used as a tool to measure quality of clinical care, particularly in low and middle income countries where medical record abstraction is unlikely to be feasible. The standardised patient method allows care to be observed without the provider's knowledge, removing concerns about the Hawthorne effect, and means that providers can be directly compared against each other. However, their undercover nature means that there are methodological and ethical challenges beyond those found in normal fieldwork. We draw on a systematic review and our own experience of implementing such studies to discuss six key steps in designing and executing SP studies in healthcare facilities, which are more complex than those in retail settings. Researchers must carefully choose the symptoms or conditions the standardised patients will present in order to minimise potential harm to fieldworkers, reduce the risk of detection and ensure that there is a meaningful measure of clinical care. They must carefully define the types of outcomes to be documented, develop the study scripts and questionnaires, and adopt an appropriate sampling strategy. Particular attention is required to ethical considerations and to assessing detection by providers. Such studies require thorough planning, piloting and training, and a dedicated and engaged field team. With sufficient effort, standardised patient studies can provide uniquely rich data, giving insights into how care is provided which is of great value to both researchers and policymakers.

Introduction

Clinical quality of care, the process through which inputs from the health system are transformed into health outcomes (Donabedian, 1988), is arguably the most informative dimension of quality, as it is the key point where provider behaviour influences case management. However, it is also highly challenging to measure (Hanefeld et al., 2017), and many commonly used methods for measuring clinical quality have significant disadvantages. Direct observation cannot control the types of patients and cases observed (Peabody et al., 2000), clinical vignettes measure knowledge rather than practice (Leonard et al., 2007, Mohanan et al., 2015), and both suffer from Hawthorne effects (Leonard and Masatu, 2010). Medical record abstraction is usually unfeasible in LMICs especially in the private sector where record keeping is often poor or non-existent (Aung et al., 2012). Patient exit interviews suffer from recall bias and poor response rates, and may require the patient to understand clinical procedures (Onishi et al., 2010).

A key advance in measurement of clinical quality is the use of standardised patients (SPs) in primary care settings. Healthy people, employed by a research study, pose as real patients, responding to the clinician's actions as a real

patient would. Alternative terms include mystery client, simulated patient, covert patient and undercover careseeker. SPs have a long history in medical education (Peabody et al., 2000), where the clinician knows that she is being tested outside a real world milieu. The method is increasingly being used as a research tool in large field studies, to assess deficits in care (Das et al., 2012, Christian et al., 2018, Kohler et al., 2017), evaluate quality improvement strategies (Das et al., 2016a, Harrison et al., 2000, Mathews et al., 2009), and identify how financial incentives influence quality (Das et al., 2016b, Currie et al., 2014).

The SP method has a number of advantages. In a high quality SP study, clinicians believe they are treating a real patient and therefore measures are not influenced by the Hawthorne effect (Leonard and Masatu, 2010). Because each case is completely standardised, care can be benchmarked against pre-determined standards for a specific condition. We can say that an antibiotic was incorrectly used because we know the SP presented with symptoms of a viral pharyngitis rather than pneumonia. The ability to control patient-mix avoids confounding and allows for the investigation of rarer conditions, such as tuberculosis (TB), which might otherwise require long observation periods to gather a sufficient sample (Peabody et al., 2000). Where the objective is to compare across different types of patients, the SP presentation can be altered (or different types of SPs such as men and women can present the same condition) to assess how provider behaviour responds to patient characteristics (Currie et al., 2011, Planas et al., 2015). Finally, in evaluations of interventions, SPs provide scope for double-blinding, whereby providers cannot tell which patients are SPs, and the SPs themselves are blinded to the treatment arm of providers they visit (Das et al., 2016a).

The main downsides are that the disease cases suitable for SPs are limited, thereby restricting their applicability, and developing SPs for use in the field is complex, which may limit their scalability. There is ongoing debate on the ethics of SP research, though the “deception” of clinicians can be ethically justified where (a) other options cannot answer the research questions (Alderman et al., 2014); (b) risks to SPs and providers are minimal; and (c) the knowledge generated is of value to society (Rhodes and Miller, 2012).

In this paper, we provide a step-by-step guide on using SPs to measure quality of care in health facilities (dispensaries, health centres or clinics). The guide is based on a review of SP studies in low and middle-income countries (LMICs) (full details in appendix), as well as our experiences implementing this approach in public and private health facilities in China, India, Kenya, South Africa and Tanzania. The SP method is also frequently used in the retail sector, for example in pharmacies or informal drug sellers (Fitzpatrick and Tumlinson, 2017), but our focus on health facilities reflects the particular challenges faced in documenting clinician-patient interactions and handling requests for exams and diagnostic tests.

Step 1: Choosing a suitable SP case

The first choice made when designing an SP study is case selection, that is, the condition or symptoms SPs present to providers. The major considerations are whether the case is technically feasible, whether it is ethically acceptable to ask SPs to present the case, and whether the case will be suitable both to the local context and the purpose of the study. We list ten questions which researchers should ask when assessing cases for inclusion in Table 1. Some cases will never be feasible and are likely to be excluded by all studies, for example, any case requiring inpatient care

would be deemed too high a risk to a fieldworker, and an SP with a wound would be practically impossible to falsify. Perceptions of feasibility may change over time; for example, TB was once perceived as a condition which could not be measured using the SP method, but has now been validated as an assessment of quality (Das et al., 2015).

It is useful to refer to – and sometimes replicate – SP cases developed by previous studies. We conducted a scoping review of all SP studies in LMIC health facilities up to December 2016, and identified 17 conditions across 63 articles, covering 45 studies (Table 2). One advantage of replicating such cases is the opportunities to share SP scripts and tools and learn from the experience of others. Colleagues can advise on the feasibility of implementing certain SP cases, and how effectively they measured quality of care. Secondly, if multiple studies share SP cases, direct comparisons are possible across settings. Examples of such comparisons to date include: (1) dispensing practices for suspected TB patients in multiple settings in urban India (Miller et al., 2018), and (2) treatment of asthma, chest pain, diarrhoea and TB across China, India and Kenya (Das et al., 2018, Daniels et al., 2017). However, as Table 2 shows, the range of SP cases used is currently limited. This may reflect the need and scope for development of more cases, but also the challenges of identifying cases meeting the requirements discussed in Table 1.

If resources allow, choosing more than one case so that each provider receives multiple visits allows more quality dimensions to be assessed and increases statistical power. One might consider using a range of different SP cases, mixing:

- Infectious diseases with non-communicable diseases (NCDs)
- Uncommon but severe conditions with common, non-critical, but high-burden diseases
- Conditions requiring lab diagnostics with those requiring only history-taking to diagnose
- Conditions for which there is typically overprovision with conditions where there is underprovision
- Different stages of disease progression or experimental variants, such as some patients already having a lab report while others do not, for the same disease

Step 2: Defining correct management

Once conditions are chosen, an indicator of correct management should be pre-defined for each SP case. Correct management should be based upon national standard treatment guidelines to ensure appropriateness to the study setting, but may need to incorporate international recommendations (such as WHO guidelines) where national guidelines are unavailable. A technical advisory group including clinicians and public health professionals, with knowledge of best practice and experience of local health systems, can also be convened to advise on correct management. Suggested types of outcomes are given in Table 3 covering both actions required, such as provision of certain drugs or referral, and actions that are not only not required, but may be considered harmful to the patient, or unnecessary care which is not dangerous but nonetheless has an opportunity cost. An alternative to a binary correct management definition is to construct a continuous index by assigning points for different elements of management. However, any such measure will be critically sensitive to the weighting of the different possible correct, incorrect, and neutral components of care. Our experience has shown that the types of unnecessary and harmful care provided can be highly unpredictable, so collecting outcomes based solely on a preconceived checklist of what *should* happen may miss much of the care that is actually provided. Researchers should therefore ensure

that data collection tools are sufficiently open and flexible to collect data on all lab tests, medicines and recommendations provided.

If the sample includes a wide range of providers or facilities, the definition of correct management may need to accommodate a range of potentially correct outcomes, depending on provider qualifications or facility level. For example, in facilities with on-site TB testing, correct management for suspected TB should be defined as the ordering of appropriate diagnostic tests. In smaller facilities without such capacity, correct management may be defined as referral to a higher-level facility.

Regardless of the provider type, researchers will need to make judgements on how lenient or strict/comprehensive the definition of correct management should be, and this can have a dramatic impact on results (Sylvia et al., 2017). Box 1 uses data from Kwan et al., 2018 to construct the flowchart of provider actions for 765 SP interactions with providers without a medical degree. If we define correct management as "asking for a TB-related test", 17.0% are classified as correctly managed. But, of these, 21.5% also gave a contraindicated drug, 42.3% did not mention TB to the patient and 30.8% gave unnecessary (but not contraindicated) drugs, including antibiotics. A stringent definition of correct management as "asked for a TB-related test without giving contraindicated or unnecessary drugs and discussed the prognosis with the patient" reduces the fraction correctly managed to 0.9%.

Further, classification of correct management may be conditional on the results of diagnostic tests. For example, correct management of suspected malaria has two steps, the second of which is conditional on the first: a malaria test must be carried out, then an appropriate antimalarial prescribed if the test is positive, or no antimalarial prescribed if the test is negative. Researchers may also wish to consider the true status of the patient in the definition of correct management. For example, if an SP is known not to have malaria, any antimalarial provision could be considered inappropriate even if the provider reports a positive test, though as such tests are not 100% accurate even under ideal conditions, this may identify both faults with the provider and with the test itself.

This complexity of defining correct management is not a flaw of the SP method per se; instead it highlights the importance of paying close attention to the definitions selected, and the utility of presenting a range of definitions. Finally, while correct management is typically the primary study outcome, it is relatively easy to also collect other outcomes related to the consultation (e.g. history taking) or the patient experience (e.g. waiting time), which provide important context for understanding correct management outcomes. Some suggestions are given in Box 2.

Step 3: Designing tools and planning the study

The SP scripts define each case in detail and are the primary means for standardising the case to ensure comparability across providers. A script begins with a short opening statement which the SP delivers to each provider describing the symptoms (such as "Doctor, I have a cough and some fever" for suspected TB), which is followed by scripted responses to history questions, which the provider may or may not ask. The SP must not give additional information to the provider outside this script, nor give information from the history question section unprompted. The script should also include a short biography describing social background, age, occupation, family details, and the circumstances of the illness presented.

The corresponding structured questionnaire, which the SP completes after each interaction, captures all information needed to define correct management (physical exams, diagnostic tests, drugs and other treatments), as well as other outcomes of interest and general comments on the visit. It should be completed soon after the visit, either as a self-completed questionnaire by the SP or through an interview of the SP by a supervisor. Developing these tools is an iterative process, and numerous changes will likely be made during piloting and training, with SP trainees themselves playing integral roles throughout this process. Steps to take when developing tools are described in Box 3.

Once the design of cases and tools are underway, the researcher must define a sampling frame and decide on the unit of analysis. Analysis of SP studies can be done at the level of the clinician or the facility. Facility-level analysis is likely to be appropriate when the research questions do not relate to the performance of specific providers, for example when evaluating an intervention randomised at the facility level. Provider-level analysis has the advantage of allowing investigators to address additional questions such as the know-do gap of individual providers (Mohanan et al., 2015), or the effect of provider cadre or training on quality. However, provider-level data are more challenging to collect because SPs must visit specific clinicians identified a priori, which presents two practical challenges: first, the production of a sampling frame of all eligible providers (facility staff lists may be incomplete and providers may work at multiple facilities) and second, the identification of providers by SPs in contexts where name badges are rare and asking for a name may be considered unusual or rude.

Step 4: Addressing ethical concerns

Ethical norms in medical research require informed, freely given consent of participants. However, the SP method, by its very nature, requires that providers do not have full information on when or how data collection occurs (Madden et al., 1997). Furthermore, because providers are likely to have substantial knowledge about the quality of their own practice, selective refusal may hamper a study's ability to produce representative data on care real patients receive (Rhodes and Miller, 2012).

Several approaches to provider consent have been used (Table 4), though it should be noted that many studies identified in the literature review (21/45) did not report their consent process.

Where consent is obtained, researchers still need to withhold certain information from participants. The participant should be given a broad window of time during which an SP will visit, not a date or appointment. For example, if SP visits are planned six weeks after consent, the provider can be informed that the visit will occur "at some point in the next three months". If the provider asks for a specific date, they should be told that to give one would compromise the nature of the research. A similar explanation should be given if they ask about the type of patient who will visit, or the condition they suffer from. To avoid providers unintentionally being given such sensitive details, ideally the team members conducting the consent process should be blinded to the SP conditions, or the consent process carried out by a senior researcher who will be able to resist pressure from providers to disclose such details. The consent process may be combined with other, non-SP aspects of a study, such as a survey of the health facility or provider knowledge.

If the waiver of consent approach is chosen, this must be justified to ethics committees, who may not be familiar with the SP method and may be wary of such waivers. Committees may only be prepared to approve such an approach if there are government approvals for the study, and/or a commitment to inform providers that they received an SP by letter or public meeting after data collection is completed. Further risks associated with using a waiver of consent are loss of the trust of a provider if an SP is discovered and risk of aggression towards that SP.

Working as an SP exposes fieldworkers to risks they would not experience during ordinary survey data collection, and it is the responsibility of the study team to minimise and mitigate these risks to the greatest possible extent. This can be achieved through two main pathways. Firstly, the study should be designed to minimise such risks. This must be considered throughout the design process, and has been discussed under other Steps, such as choosing SP conditions that minimise the risk of fieldworkers undergoing invasive tests. Secondly, fieldworkers should be trained intensively to avoid risks which cannot be removed by design (Table 5). One risk-minimising strategy SPs will frequently need to use is the refusal of invasive tests; a particular challenge is ensuring that the reasons given for refusals come across as normal behaviour and do not raise suspicions. Despite these challenges, experience has shown that the SP method has minimal risk to fieldworkers equipped with proper training (Daniels et al., 2017) and need not inconvenience real patients (Das et al., 2015).

Step 5: Training fieldworkers and organising fieldwork

Playing the role of an SP is more complex and demanding than standard fieldwork, so we recommend recruiting experienced and proven fieldworkers. While some studies have recruited trained actors, experience indicates that while actors may perform well in improvisation and staying in character, adherence to protocol and precise recall of information are equally important. Many studies have therefore drawn from the same population they would use for any survey enumerator position and dedicated several weeks to selecting and training on SP skills.

The mix of SPs may also matter if quality is expected to vary by age, social group, or other characteristics. For example, male and female SPs may receive different treatment (Borkhoff et al., 2009), so for cases relevant to both genders, hiring an even mix of men and women and randomly assigning them to visits should be considered. Alternatively, cases may be portrayed by one gender only; this may be appropriate for cases such as family planning clients, but for other conditions may make the study less generalisable. Researchers should consider whether SPs will need a certain physical appearance to portray the case (for example, a 60-year-old woman could not portray a family planning client), and the languages spoken by typical patients in the geographical areas of interest.

Administering a background health questionnaire at the start of training is a crucial first step for protecting fieldworkers, maintaining consistency of SP case presentation, and ensuring that real health conditions do not confound the interpretation of results. For example, the physical symptoms of poorly controlled asthma or hypertension may lead a provider to dismiss a possible diagnosis of TB in an SP with a cough and chest pain. This may require consultation with your institution's Human Resources department to check that equal opportunity requirements are balanced with study needs.

Training should begin with an introduction to the concept of SPs, followed by fieldworkers reading and role-playing scripts. They should work in small groups to discuss the patient narrative and identify difficulties with phrasing or context-specific inconsistencies. For example, in a Tanzanian training session run by some of the authors, an initial draft of a script instructed the SP to say that they had never had an HIV test, but trainees noted that this would be implausible for female SPs with children, since HIV testing is ubiquitous in antenatal care there.

Emphasis should be placed on playing the role consistently, never giving more initial information than the opening statement, and then providing answers to only the questions the provider asks, which is essential for ensuring measurement reliability. As they learn about the study condition it can be tempting for SPs to help or guide the provider to a correct diagnosis, so training must explain why it is important to avoid this. Comparison across SP studies has confirmed that the amount of information provided heavily influences treatment choices by providers (Miller et al., 2018).

In most studies, each fieldworker performs only one SP case throughout the study. However, training fieldworkers in two roles gives the team more flexibility, though SPs should be randomly allocated to a role at each facility to avoid bias. In studies covering large geographies, it may not be possible for SPs to be randomly allocated to facilities, and an SP-specific variable should be controlled for as a fixed effect in analysis (Das et al., 2016a). There should be no systematic differences in time of day or week of the visit by condition or SP – for example, avoid the male SPs always visiting in the morning and female in the afternoon.

In studies in rural or remote locations, particular attention should be paid to 'cover stories', or how SPs explain their presence as an outsider if questioned. One resource-intensive approach is to research in advance the names of villages and people who SPs can say they are visiting, specific to every location. Alternatively, a number of stories can be developed for use in different contexts: e.g. that they are buying cash crops or livestock or researching places to sell second-hand clothing. Experience in the field has taught us that SPs should not improvise: some members of a team were detected after telling one provider they were agents for the government.

Once SPs understand their script and role, introduce them to the questionnaire. A useful training exercise is to have fieldworkers observe the same role-play, then complete the questionnaire separately. Comparing answers highlights difficult parts of the consultation to remember. The final stage of training is SPs practising their roles and questionnaires by making undercover visits to providers who have agreed to take part. It may be helpful for this to initially be done in pairs (e.g. posing as husband and wife) so that peer feedback can be provided.

If SPs are permitted to undergo certain diagnostic tests (for example, fingerprick blood tests or urinalysis), we recommend that supervisors retest any fieldworker who receives a positive result for malaria or urinary tract infection. This will give peace of mind to the fieldworker (or allow for treatment if a true positive) and validate the facility's test for the purpose of analysis. Supervisors can be trained to conduct mRDTs and urine dipstick tests and be provided with a supply for the field.

SPs should purchase all drugs prescribed, if the budget allows, as this will reduce recall bias when recording drugs prescribed, improve the comprehensiveness of data on medicines, allow for collection of drug costs, and reduce the risk of raising provider suspicion. Additionally, it may be possible to incorporate drug quality testing into the study

(Wafula et al., 2017). To test reliability of recall, SPs can carry covert audio recorders, although this may introduce additional ethical issues (Das et al., 2015).

Step 6: Assessing detection

A follow-up study to assess the detection rate of SPs (that is, the proportion of SPs identified by providers as being SPs and not genuine patients) is seen as an important step in ensuring the validity of results. Detection rates from recent health facility LMIC studies have typically varied from 0 to 5% (Sylvia et al., 2017, Daniels et al., 2017, Das et al., 2015), but there is no consensus on a maximum acceptable rate. Higher detection rates can be expected in rural settings compared to urban ones, where outsiders are likely to raise more suspicion. False positive rates (providers report suspecting real patients to be SPs) varied from 1 to 6% in the same studies.

It may be advantageous to inform providers when obtaining consent that there will be a follow-up study and ask them to make a note of the name, description, symptoms and date if they receive any patients they suspect are SPs. This will allow for easy distinction between true and false detections at follow-up. However, priming providers in this way may increase the risk of detection, so the study team must decide whether they are willing to take this risk for the benefit of ease of classification. In addition, priming is not possible where a waiver of consent or institutional consent is used.

Dependent on setting and resources, the detection survey can be conducted as a face-to-face interview, or remotely by telephone or email. If face-to-face, the survey can be combined with other elements of the study, such as vignettes to measure provider knowledge and compare with SP performance to measure the know-do gap (Das et al., 2015, Sylvia et al., 2017, Mohanan et al., 2015). Carrying out such knowledge assessments after completion of SP visits has the advantage of being less likely to influence provider behaviour than if done before SP visits. In addition, if a waiver of consent has been used, the detection survey is an opportunity to inform providers that SP visits have taken place and allow them to ask questions and provide feedback.

The detection survey should start by briefly reminding (or in the case of a waiver of consent, informing) providers of the SP study's aims and methods, then asking if the provider recalls receiving patients they suspected were SPs. For every suspected SP, the following information should be collected:

- Date and time of visit (approximate if necessary)
- Name, (approximate) age and gender of SP
- Symptoms of SP
- Diagnosis and treatment given by provider
- The reason the provider suspected the patient was an SP
- Whether the provider became suspicious during the visit or after it was complete
- Whether the provider changed their treatment or confronted the SP due to their suspicions

These data should then be used to classify suspected SPs as true or false positives at the analysis stage. The stringency of a true positive definition will depend on setting, conditions and whether providers are primed. Some

studies may require that the name of the SP is reported, but others may only require that the provider correctly identifies the gender and symptoms of the SP and gives a date of visit correct to within one week.

Conclusion

SPs are a valuable research tool, with enormous potential to improve the measurement of clinical quality in primary care settings. However, their undercover nature means that there are methodological and ethical challenges beyond those found in normal fieldwork. Moreover, SPs in health facilities are much more complex to implement than those in retail outlets. There is growing experience of developing and implementing a range of SP cases in diverse settings, and we hope that this paper can help make such learning accessible to those planning similar studies.

The choices made when undertaking an SP study are highly dependent on the setting, purpose and resources. A well-designed study will draw on a thorough understanding of the health system in question. It will also capitalise on the contribution of fieldworkers during tool development, training and piloting to ensure cases are credible, rarely detected, and minimise risk. The task of developing the script, backstory, symptoms and behaviour of an SP should not be underestimated. The process of implementing SPs must therefore be collaborative, incorporating both local knowledge and technical expertise on the SP method.

The absence of Hawthorne effects and the ability to observe healthcare as it is delivered, while controlling the condition and characteristics of that patient, make SPs a valuable tool which can answer research questions no other method can. We also recognise that the SP method, as currently implemented, has its limitations. With this in mind, we conclude by offering a number of avenues for future methodological development (Box 4). These relate to challenges in investigating continuity of care, defining correct treatment in different contexts, dealing with false positive diagnostic tests, conducting power calculations and representativeness of the population of patients.