


RESEARCH ARTICLE

Open Access



# Development of a video-observation method for examining doctors' clinical and interpersonal skills in a hospital outpatient clinic in Ibadan, Oyo State, Nigeria

Dr Navneet Aujla<sup>1\*</sup> , Temitope Illori<sup>2,3</sup>, Achiaka Irabor<sup>2</sup>, Abimbola Obimakinde<sup>2,3</sup>, Eme Owoaje<sup>3</sup>, Olufunke Fayehun<sup>3</sup>, Motunrayo M. Ajisola<sup>3</sup>, Sinmisola O. Bolaji<sup>3</sup>, Samuel I. Watson<sup>4</sup>, Timothy P. Hofer<sup>5</sup>, Akinyinka Omigbodun<sup>3</sup> and Richard J. Lilford<sup>4</sup>

## Abstract

**Background:** Improving the quality of primary healthcare provision is a key goal in low-and middle-income countries (LMICs). However, to develop effective quality improvement interventions, we first need to be able to accurately measure the quality of care. The methods most commonly used to measure the technical quality of care all have some key limitations in LMICs settings. Video-observation is appealing but has not yet been used in this context. We examine preliminary feasibility and acceptability of video-observation for assessing physician quality in a hospital outpatients' department in Nigeria. We also develop measurement procedures and examine measurement characteristics.

**Methods:** Cross-sectional study at a large tertiary care hospital in Ibadan, Nigeria. Consecutive physician-patient consultations with adults and children under five seeking outpatient care were video-recorded. We also conducted brief interviews with participating physicians to gain feedback on our approach. Video-recordings were double-coded by two medically trained researchers, independent of the study team and each other, using an explicit checklist of key processes of care that we developed, from which we derived a process quality score. We also elicited a global quality rating from reviewers.

**Results:** We analysed 142 physician-patient consultations. The median process score given by both coders was 100%. The modal overall rating category was 'above standard' (or 4 on a scale of 1–5). Coders agreed on which rating to assign only 44% of the time (weighted Cohen's kappa = 0.26). We found in three-level hierarchical modelling that the majority of variance in process scores was explained by coder disagreement. A very high correlation of 0.90 was found between the global quality rating and process quality score across all encounters. Participating physicians liked our approach, despite initial reservations about being observed.

(Continued on next page)

\* Correspondence: [NAujla@warwick.ac.uk](mailto:NAujla@warwick.ac.uk)

<sup>1</sup>Warwick Medical School, University of Warwick, C/O Room B147a, CV4 7AL  
Coventry, United Kingdom

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(Continued from previous page)

**Conclusions:** Video-observation is feasible and acceptable in this setting, and the quality of consultations was high. However, we found that rater agreement is low but comparable to other modalities that involve expert clinician judgements about quality of care including in-person direct observation and case note review. We suggest ways to improve scoring consistency including careful rater selection and improved design of the measurement procedure for the process score.

**Keywords:** Quality of healthcare, consultation quality, low-and middle-income countries, ambulatory care, physicians, video-observation

## Background

Improving the quality of primary healthcare provision is now an imperative for low-and middle-income countries (LMICs) [1–6]. Recently published reports suggest that health systems in LMICs face several challenges, including workforce shortages, and limited supplies of medication and necessary medical equipment [6–8]. There are also problems with the quality of care delivered in individual healthcare encounters in LMICs, including incorrect diagnoses; poor adherence to clinical guidelines (< 50 % on average); medication errors; and provision of inappropriate care, such as unnecessary surgical interventions [6]. In addition, users express concerns about a lack of compassionate and respectful care, and low empathy [6].

In order to develop effective interventions to improve the quality of care, we first need to be able to accurately measure quality in order to better understand where any problems lie and to evaluate the effects of an intervention [9]. This paper focuses on the ‘process’ component of quality of care [10], which describes the clinical processes involving patients, such as the action of ordering a test or conducting an examination when necessary [11]. Donabedian [12] referred to this as the technical quality of care.

The most commonly used measurement methods for technical quality of care all have some key limitations in LMICs settings. Medical record documentation is sparse [13–15], in-person methods can be expensive and cumbersome to arrange [16] and, as Miller [17] suggests, testing providers may not measure practical performance in daily clinical practice. There is also the problem of the Hawthorne effect, which describes a change in behaviour as a result of being observed [18], and is evident when an observer is physically stationed in the consultation room [19]. Video-observation has a number of appealing features and offers an attractive combination of a comprehensive capture of events during an encounter, relatively low impact on the efficiency of routine healthcare delivery, and the ability to carry out the assignment of raters to encounters in a flexible and efficient way. We describe preliminary feasibility and acceptability data on the use of video in assessing the quality of clinical

encounters in a LMIC setting. We also develop a checklist to evaluate consultation quality and report on the measurement characteristics of this tool. Measurement checklists that have been used to evaluate key general and symptom-specific processes of care in existing studies in LMICs tend to focus on child illnesses and have been designed for in-person observation [20–24], so are often too long and impractical to use to efficiently code video-observed consultations.

We tested our approach in the outpatient department of a large tertiary care teaching hospital in Nigeria, based on our existing work on the NIHR Global Health Research Unit on Improving Health in Slums project (<https://warwick.ac.uk/fac/sci/med/about/centres/cahrd/slums/>). The Improving Health in Slums project examines healthcare access and use by slum dwellers across multiple sites in South-Asia and Sub-Saharan Africa. We have evidence that a substantial proportion of people in cities seek doctor and nurse outpatient consultations at outpatient departments in hospitals. The technical quality of care provision by individual providers in the community (such as pharmacies and single-handed practices) in LMICs is reported to be poor in many studies [6, 25–30], with evidence of practices such as prescribing antibiotics for unstable angina [31]. However, only a few studies have been carried out to assess the technical quality of care in a hospital outpatient setting in LMICs [32–37]. These mostly provide only patients’ views and to our knowledge, none have used video-observation to measure the quality of care, despite the promise of the approach.

Therefore, the aim of our study is to conduct a preliminary examination of feasibility and acceptability of video-observation for assessing technical quality of care provision by doctors in a hospital outpatients’ department in a LMIC setting. A further aim is to develop measurement procedures and gain information about measurement characteristics, including validity and reliability of the measurement. Our goal was to prepare the ground for our future work on the assessment of quality of care in LMICs and to identify how we can further develop and refine our approach and measurement procedures.

## Methods

### Design and setting

We conducted a cross-sectional study in the general outpatient department of University College Hospital (UCH) in Ibadan, Oyo State, Nigeria. The UCH is a 1,000 bed teaching hospital located in south-western Nigeria. The general outpatient clinic of the hospital is the entry point for most patients presenting to UCH for primary health care services. The services are provided by consultant family physicians and supervised family physicians in training.

The study involved undertaking video-recorded observations of consecutive physician-patient consultations, followed by brief semi-structured interviews with participating physicians to understand their perceptions and experience of being video-taped during the consultations.

### Participants and eligibility criteria

The unit of observation was a single encounter with an adult patient (or child and parent/caregiver). Participating physicians were resident doctors consulting at the clinic at the time of the study who provided written informed consent to take part. Adult (> 18 years) and child patients (under 5 years) consulting with a participating physician were then also eligible to take part. Written informed consent was sought from adult patients. For child patients, written consent was provided by parents/caregivers. We did not exclude participants on the basis of presenting complaint. However, patients were only eligible if they were consulting for a new problem. Patients that did not live within Ibadan City – where UCH is situated – were excluded since our focus was care for local residents.

### Sampling and recruitment

Eligible physicians and patients were sampled opportunistically from the triage clinic, based on attendance on the day of each clinic. An average clinic week runs from Monday to Friday with two clinics (morning, 8am-1pm and afternoon, 2pm-6pm) each day. We were present in the clinic for all sessions over 10 consecutive working days from 1st to 12th April 2019. Eligible patients were identified and initially approached by a clinic coordinator. Two local project researchers were stationed at the clinic to recruit and consent individuals interested in taking part. Yoruba translated versions of the participant information leaflets and consent forms were available to patients as necessary. Ethical approval was granted by the University of Warwick Biomedical and Scientific Research Ethics Committee (REGO-2018-2306) and the University of Ibadan/UCH, Ibadan Research Ethics Committee (UI/EC/18/0646).

### Sample size

Between 60 and 70 adults and children under 5 are estimated to present on a regular clinic day at the Family Medicine Department at UCH. The number of consultations we could examine was limited by resources, but we aimed to recruit at least 120. A sample of this size would provide precision enough to estimate the mean percentage quality score in the population with a 95 % confidence interval of  $\pm 3.7$  % points at a maximum (at a value of 50 %).

### Procedure

The video-cameras were stationed in two designated consultation rooms in close proximity to the clinic's waiting area. Two physicians were video-recorded simultaneously. Each video-camera was carefully positioned to ensure an unobstructed view of the physician. We ensured that the patient's face was not in view and as little as possible of the back of their head was captured on the video-recording. The video-cameras were managed by the study researchers and a technician who helped to manually start and stop the recording as a patient entered and left the consultation room. Our study procedures were first piloted in the clinic. The main sample of video-recordings were later double-coded by two medically trained researchers, independent of the study team and each other, using a specially designed checklist – details for which are provided below. The coders were trained before video-coding commenced.

The study researchers carried out brief interviews with participating physicians after they had finished video-recording the full set of consultations for each physician. The brief interviews were semi-structured and facilitated through use of a topic guide that covered physicians' reactions to being video-recorded as part of this study and potentially in future research and their prior experience of being involved in video-recorded observations (such as in medical training) (see [Appendix](#)). The interviews were conducted in-person and lasted around 10 min. A written record of the conversations was captured by the researchers in note-form. The notes were typed up, translated where necessary and securely shared as Word documents for analysis.

### Outcomes and measures

We examined four tracer symptoms: fever, cough, diarrhoea, and abdominal pain. We chose these symptoms because they are common in many LMICs, in order to enhance the generalisability of our work, and these symptoms are also red-flags for serious conditions including malaria, diarrhoea and tuberculosis. Consultations covered patients with these symptoms and patients who did not have these symptoms.

We used two approaches to measure consultation quality: an explicit checklist of key processes of care and a single global judgement-based question. Both measures examine technical and interpersonal skills including empathy. The criteria on the checklist were grouped according to the main components of the clinical encounter (interviewing/history-taking, physical examination, diagnosis and treatment, and counselling) [38], and applied to adults with specific items for child patients (see Table 1). All general criteria were applied to each consultation, and were developed based on expert feedback and adapted from criteria on checklists used in prior studies [20–24, 39]. We used these existing studies to also adapt and develop new criteria for symptom-specific clinical management alongside relevant clinical guidelines, such as the Standard Treatment Guidelines for Nigeria [40], World Health Organisation (WHO) guidelines on Integrated Management of Childhood Illnesses (IMCI) [41] and Integrated Management of Adult and Adolescent Illness (IMAI) [42]. We established a pool of explicit process measures suitable for evaluation with video-observation. The pool of criteria was subsequently reviewed by local and international clinical experts from the research team to ensure content validity and was further modified on the basis of their feedback. Raters used their own clinical judgement to guide the selection of relevant criteria in each symptom-specific checklist module.

The single global judgement-based question we used was adapted from Rubenstein [43]: Considering everything you have seen of this encounter, how would rate the overall quality of care delivered to this patient? It was developed for estimates of quality of care based on medical record review to provide an overall impression of observed quality of care in each consultation. Responses were made on a five-point Likert scale as follows: well above standard, above standard, adequate, below standard, well below standard. We presumed that this judgement-based measure would have lower reliability (or precision) than an explicit measure based on checklists. However, by virtue of allowing an expert rater to take into account a wide variety of relevant information and context apparent in the video but not captured by the explicit checklist, it has appealing strengths in terms of the validity of measurement that are distinct from but competitive with the validity conferred by the expert panels commonly used to develop explicit checklists based on guidelines. In many studies explicit and implicit judgement measures have been compared as a way to provide convergent validity for the use of both to assess quality of care and we included this measure for this purpose [44–48].

Using the rater responses, we derived two measures of quality:

- ‘Process quality score:’ the process quality score was derived using general and symptom-specific responses on the checklist. Process scores were calculated for each physician by dividing the number of positively identified criteria (numerator) by the total number of checklist criteria (general and symptom-specific) that applied for that consultation (denominator). This followed similar approaches used elsewhere (RAND Health, [https://www.rand.org/health/surveys\\_tools/qatools.html](https://www.rand.org/health/surveys_tools/qatools.html)).
- ‘Global quality rating:’ we established the global quality rating based on responses to the judgement-based question. Each physician rater completed a global quality rating for each assigned encounter.

Every consultation was assigned a process quality score and a global quality rating. However, process quality scores for consultations involving patients that had one or more tracer symptoms were derived from assessments of both general and symptom-specific criteria, and assessment of only the general criteria for consultations involving patients without any of the four symptoms (see Fig. 1 in the next section).

### Analysis methods

As process quality scores were not normally distributed and limited to [0 %, 100 %], we describe the median and interquartile range, and report the mode for the global quality rating. These analyses were conducted for each coder and reported separately for adult and child patients.

We examined measurement characteristics of the checklist using the following approach. For process quality scores, we estimated a Bayesian hierarchical model with three-levels including physician, patient encounters within physicians and rating occasions within patient. The model included no explanatory covariates. We estimated the proportion of total variance in quality scores that was attributable to the treating physician, differences in the ‘true’ quality of care received by one patient and the variation between rating occasions [49]. In this model, ‘true’ quality of care represents the quality score that would be obtained by an average over a very large number of rating occasions and treating physicians for an individual patient encounter. Weakly informative prior distributions were specified. The distributions used were the half t-distribution with 4 degrees of freedom ( $t_4$ ) for hierarchical standard deviation terms and the normal distribution with mean 0 and standard deviation of 5 ( $N(0,5^2)$ ) for model coefficients. These variance

**Table 1** Assessment criteria, based on existing literature [20–24, 39–42] and expert feedback

General
<ul style="list-style-type: none"> <li>• Greeted patient/carer</li> <li>• Solicits what the problem is and allows patient to fully elaborate presenting problem</li> <li>• Exhibits well organised approach to information-gathering</li> <li>• Gave due attention to patient/carer (looking and listening)</li> <li>• Washed hands</li> <li>• Number of minutes spent examining patient behind the screen</li> <li>• Arranges appropriate follow-up</li> <li>• Gives patient a clear explanation of the condition, the treatment, what to look out for.</li> </ul>
Cough symptom
<ul style="list-style-type: none"> <li>• Asked duration of cough</li> <li>• Asked about difficulty in breathing</li> <li>• Asked about wheezing</li> <li>• Asked about presence of fever</li> <li>• Asked about sputum production</li> <li>• Asked about TB history and exposure</li> <li>• Listened to lung</li> <li>• Told to return quickly if: breathing becomes difficult, child unable to drink, child becomes more ill, child has convulsions</li> </ul>
Fever symptom
<ul style="list-style-type: none"> <li>• Asked about duration of fever</li> <li>• Asked about localising symptoms suggesting site of infection if not obvious (headache, neck stiffness, skin, mouth and pharynx, lungs, urinary tract, gastrointestinal tract)</li> <li>• Site of infection obvious (Yes/No)</li> <li>• If yes, examined for localising symptoms if site not obvious (neck stiffness, skin, mouth&amp;pharynx, lungs, urinary tract, GI tract)</li> <li>• (If infant with high fever) gave paracetamol/aspirin in correct dosage</li> <li>• Advised increased fluid intake</li> <li>• Told to return in 3 days if fever persists</li> </ul>
Diarrhoea symptom
<ul style="list-style-type: none"> <li>• Asked duration of diarrhoea</li> <li>• Asked about presence of blood or mucus in stools</li> <li>• Asked about vomiting</li> <li>• Asked about HIV status/CD4 count</li> </ul>
Checked for dehydration:
<ul style="list-style-type: none"> <li>• Checked abdomen</li> <li>• Pinched skin examining for signs of severe dehydration</li> <li>• (If infant) checked for sunken fontanel</li> <li>• Treated dehydration appropriately</li> <li>• Referred case if severe or blood in stool</li> <li>• Kept child under observation if moderately dehydrated</li> <li>• Advised increased fluid intake until diarrhoea stops</li> <li>• Told how to prepare and administer oral rehydration solution</li> </ul>

**Table 1** Assessment criteria, based on existing literature [20–24, 39–42] and expert feedback (*Continued*)

<ul style="list-style-type: none"> <li>• Told to return in 3 days if child does not improve or quickly if danger signs of dehydration appear</li> </ul>
Abdominal pain symptom
<ul style="list-style-type: none"> <li>• Asked about duration and progression</li> <li>• Asked about presence of fever</li> <li>• Asked about weight loss and appetite change</li> <li>• Asked about blood or mucus in stools</li> <li>• If female, asked about last menstrual period; chance of pregnancy</li> <li>• Examined abdomen for location and nature of pain, and distension</li> <li>• (If acute abdominal pain) checked for rebound tenderness</li> <li>• Told to return if: pain worsens, unable to tolerate liquids without vomiting, fever present</li> </ul>

components were used to calculate reliability coefficients for the measurement of quality.

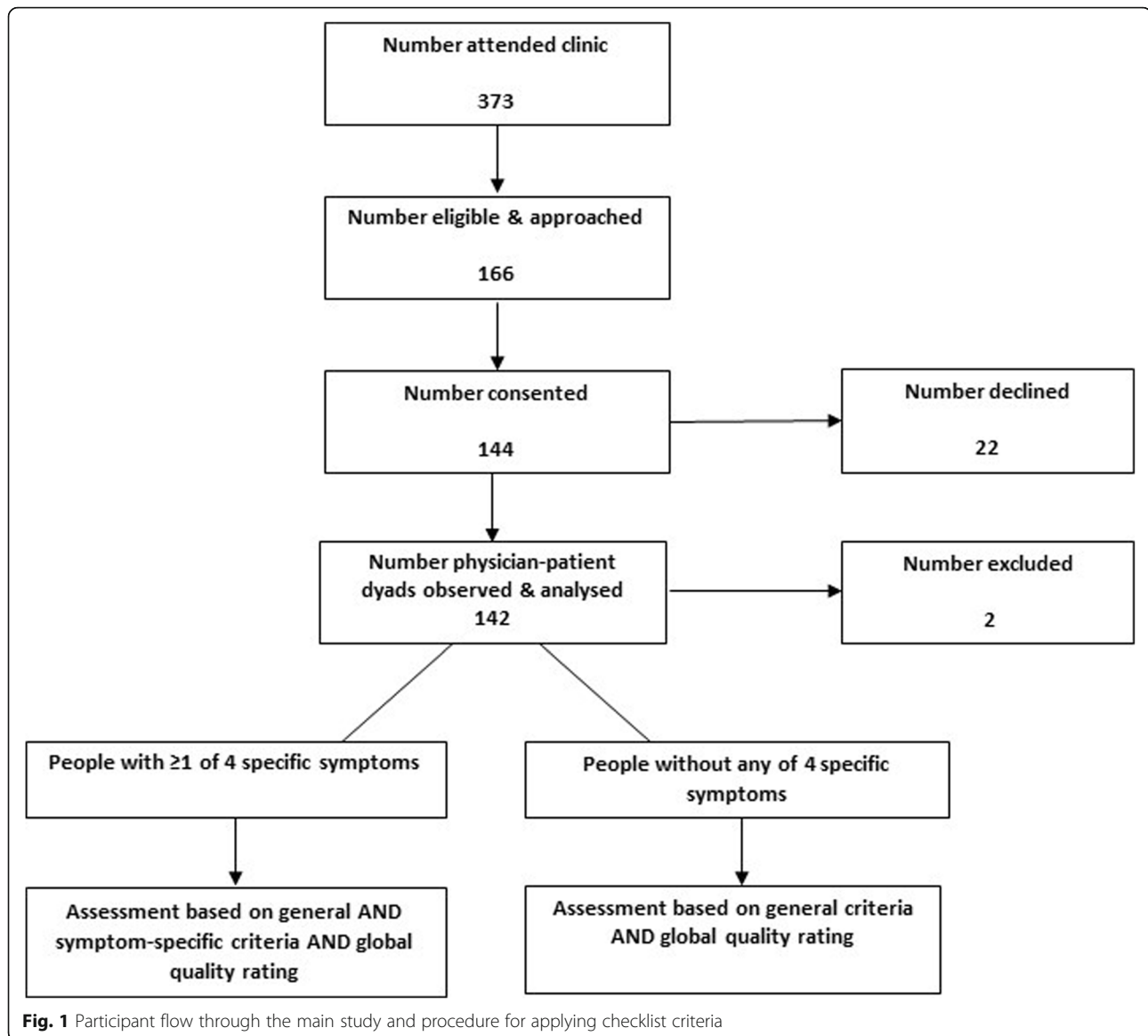
For the global quality rating, we describe agreement between coders visually using scatterplots, and by calculating a weighted Cohen's kappa, which assumes a simpler model than the hierarchical model above. In order to quantify the correlation of the process score and the overall quality rating, we estimated a joint hierarchical model with the global quality rating and process quality score as outcomes, adjusting for coder effects and adult/child differences, and included a bivariate normally distributed random effect for patient in the models. Quantitative data analyses were undertaken using R version 3.4.4 and the hierarchical model was estimated using Stan 2.19.

The analysis of the brief interviews with participating physicians was guided by a thematic approach [50, 51], which was adapted to suit our data. The analysis was carried out by NA, in consultation with the interviewers (MMA and SOB).

## Results

### Participants and feasibility of the method

Nine eligible physicians were consented to the study. None of the physicians we approached declined to participate. Five out of the nine physicians were male (56%). Most of the physicians (67%) qualified in 2007 or after and had been practising for a median of 12-years (interquartile range (IQR): 10, 13). Figure 1 shows the flow of patient participants through the main study. Of 373 patients who attended the clinic during the recruitment period, 166 were eligible and they were invited to take part. The remainder were ineligible to take part because they resided outside of Ibadan City (> 95%) or were involved in the piloting phase of the study. Twenty-two patients declined at this stage for several reasons including: waiting time in the clinic and a desire not to extend this time to participate in the study, a



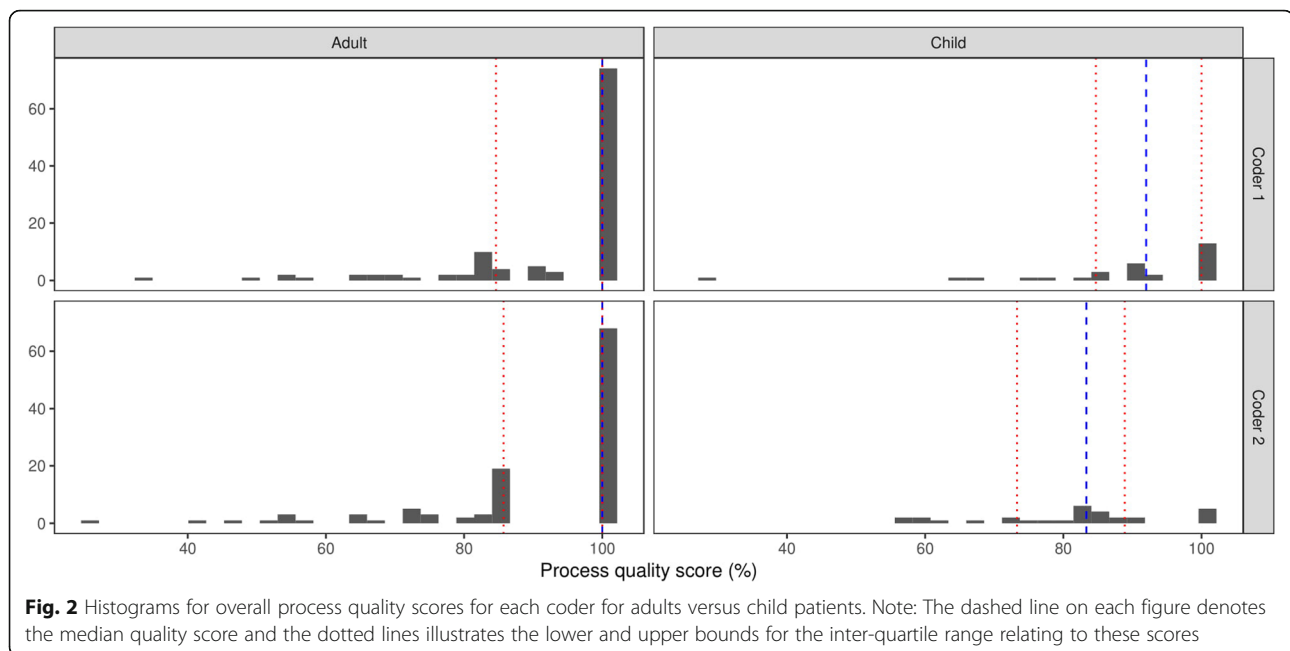
general lack of interest in taking part in the study and a preference not to have their consultation observed. Overall, 144 patients gave their consent and were observed. Two patients were excluded from the analysis because a technical issue resulted in no video-recording for these patients. We analysed 142 consultations— 112 adults and 30 children under 5 – and a process quality score and global quality rating were applied to all of these.

#### Measurement characteristics of the process checklist

The median process quality score (calculated based on both general and symptom-specific criteria) given by both coders was 100%. Figure 2 shows that the majority of process scores were 100% (coder 1: IQR: 85–100%, coder 2: IQR: 83–100%). The modal category on the

global quality judgement-based question was ‘above standard’ (see Fig. 3). Coder 1 rated 79% of consultations as above or well above standard and 93% for coder 2. These findings were consistent across adult and child consultations. Process quality scores were consistently high for all physicians. While the number of consultations observed varied across physicians across these consultations, seven out of the nine physicians included in the study achieved a median process score of 100% (see Appendix Table A). The lowest score for the remainder was 83%.

The coders agreed on process quality scores in 50 (35%) of the 142 consultations (see Appendix Figure A). Most of the general criteria (six out of seven) were used



by each coder in each consultation. For the symptom-specific modules of the checklist, coders chose how many of the symptom-specific criteria to apply in each consultation. Therefore, not all of the criteria were necessarily used for each presented symptom. Each coder included at least one criterion from each module. The overall median number of criteria used by each coder per rating occasion was six criteria. The more criteria chosen in each consultation, the worse the resulting process quality score seemed to be for the consultation (see Appendix Figure B).

The hierarchical model estimate for the process quality score was that the treating physician only accounted for 2% of the variance. This indicates little difference in average process score between physicians. Differences in the quality of care received by patients within physician (i.e., between different consultations administered by the same physician) accounted for 22% of the variance. The remainder of the variance was explained by variation in scores between rating occasions for the same patient encounter. For our purposes this variance component represents the noise in our measurement of quality of care. This suggests that the reliability of using the process score to distinguish between patients would be 0.24 if you selected randomly selected patients (with random providers caring for them) and measured with a single randomly selected reviewer.

#### Measurement characteristics of the global quality rating

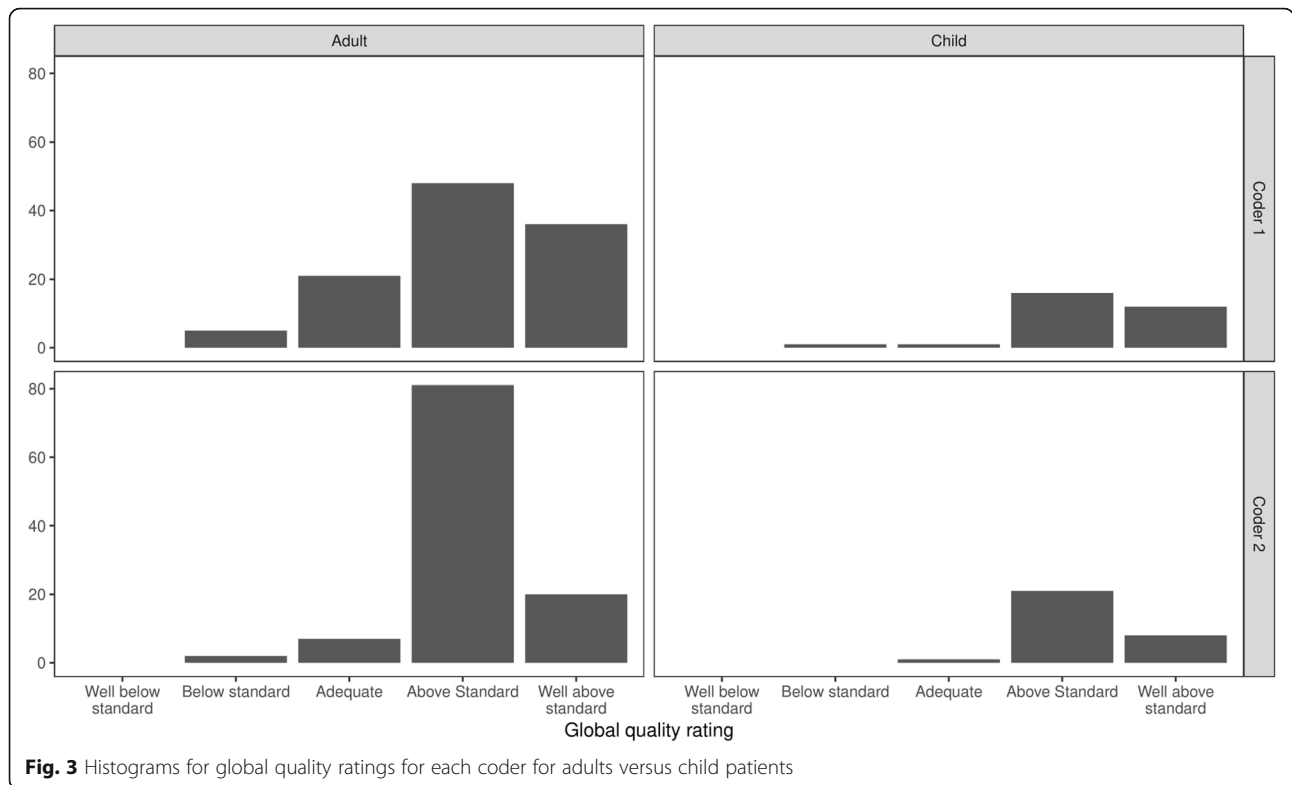
The global quality rating represents a one question summary judgement as opposed to the multi-component

process score reported above. Overall, for the global quality rating, the coders agreed on which rating to assign only 44% of the time, if exact agreement is required for 'agreement.' Requiring exact agreement does not make allowances for ratings that are only one category off but not exactly the same. The kappa statistic with quadratic weights gives some credit for agreement to scores that are close to each other but not exactly the same. The resulting statistic represents the reliability of the measurement for distinguishing between patients. It is comparable to an intra-class correlation reliability coefficient [52] and to the reliability calculated above for the process score. The weighted kappa was 0.26, which is described as representing 'fair' agreement [53].

The process score based on explicit checklists of processes of care deemed to represent good quality care and the reviewer's implicit global judgement of quality have been used in prior literature [54] as two ways of measuring the same concept – encounter quality of care. The estimated correlation of the underlying 'true' quality of care (as defined above in the methods) for the two measures across all encounters in our study was very high at 0.90, as estimated with a joint hierarchical model that removes the measurement noise.

#### Physicians' views on acceptability

We received brief feedback about our approach from all nine participating doctors. Three basic themes emerged from the data, which we briefly summarise below and indicate the number of physicians that cited each feature of each theme.



**Performance monitoring**

Most of the physicians (six out of nine) had no previous experience of being video-recorded while consulting with patients. Five out of nine reported that they liked the approach. Two out of nine explicitly stated that they would be content to be videoed again. One physician said that providing feedback to providers on their performance during the consultation should be incorporated into any performance reviews that they receive as part of their roles.

**Awareness of the video-camera**

Although around one-third of physicians said that it felt unnatural at first to be observed by the video-camera, they perceived that they found it easy to habituate and consult as they usually would. Four of the nine physicians reported that they ignored the camera from the outset of the consultation and a further two said that they forgot it was there. One physician reported that the presence of the video-camera may have encouraged them to perform better during the consultation.

**Practical improvements to the approach**

One of the physicians that reported initial self-consciousness also said that the position of the camera was too obvious. This was also reported by one of the physicians that ignored the camera. Two out of the nine physicians found it distracting for the video-camera to

be manually switched on and off between consultations. Five of the nine physicians recommended inconspicuous placement of the camera so that both the doctor and patient would find it less distracting. Two participating physicians said that informed consent should be taken on a different day to video-recorded observations to help minimise their sensitisation to being observed.

**Discussion**

There is a need to develop methods to accurately measure the technical quality of primary healthcare provision in a hospital outpatient department setting in LMICs. We developed a video-observation method to address this need and sought to assess the feasibility and acceptability of this mode of measurement. We were able to video-observe 142 doctor-patient consultations in the outpatients’ department of a large, tertiary care hospital (UCH) in Ibadan, Nigeria, with minimal disruption to the clinic’s daily work. Physicians were willing to participate and many told us that they liked our approach, although around a third said that they had some initial reservations about being observed.

In interviews with the physicians in our study they acknowledged that they thought about the camera suggesting the possibility of a Hawthorne effect [18], the magnitude of which we are not able to discern from our work. Prior studies of in-person direct observation in LMICs suggest that the magnitude of the Hawthorne



effect from an observer in the room may be small [24] to moderate [16, 19], but declines with greater numbers of observations. Extended observation periods where possible would give participants time to desensitise to the presence of the video-camera [55]. Participants in our study also made some practical suggestions such as taking informed consent on a separate day to the observations; inconspicuous placement of the video-camera; and less intrusive alternatives to our process requiring the researcher/technician to manually start and stop the video-recording as a patient entered and left the consultation room.

One physician suggested to provide performance feedback to healthcare workers based on the observed consultation. It is possible that the opportunity to get feedback after measurement may enhance the perceived acceptability of observation to healthcare workers [56]. Further examinations of the acceptability of methods of assessing quality of care in LMICs are required [57], from the perspective of both healthcare workers and patients, including those who do and those who do not participate in studies using the approach.

An important, if familiar, issue that emerged in our study relates to the relatively low reliability or lack of precision of the measurements of quality of care. The weighted Cohen's kappa ( $\kappa = 0.26$ ) relating to coder agreement in assigning a single overall global quality rating to an encounter demonstrates only fair agreement. However, it is entirely consistent with the 0.2–0.4 range cited in the existing literature for implicit expert review of quality of care by experts using medical record review [58]. Given the overwhelming stability of this estimate in the literature it is unlikely that it can be improved much. At this level of reliability for a single rating, 12 independent reviews would need to be averaged to achieve a reliability of 0.8 to distinguish between individual cases [59]. The number of reviews needed to distinguish between sites of care could be more or less depending on the magnitude of the differences in quality of care between sites relative to the quality differences between patients within sites.

Given that the process quality score is based on more explicit process criteria, we had hoped that it would provide more consistent measurements with higher reliability. Yet, we observed a similar reliability of 0.24 for this measurement of quality, although reassuringly the two measures appear to be measuring the same latent construct representing quality of care as evidenced by the high adjusted correlation between the global measure and the process score. Given how much more difficult it is to curate and maintain the process quality measurement, this raises the question of whether a much simpler rating based on global judgement might be used.

However, the more detailed information available from the process scores is alluring and there are several ways that this pilot work suggests that reliability of the video process scoring could be improved.

First, the overall number of criteria used was relatively small for any encounter. While the symptom probes were designed to cover common presentations, the match was clearly not ideal for the studied population. There were very few symptom-specific consultations to code overall, so more of the general criteria tended to be applied when evaluating the consultations (see Table 1). Furthermore, coders were left to independently decide when to use the symptom-specific criteria and how many were applicable to the consultation in question. In consultations where symptom-specific criteria were deemed to be relevant, coders only applied a median of one criterion. Therefore, symptom-specific coding was done using individual, rater selected criteria rather than the entire group of criteria available on the checklist, thus increasing rating occasion sources of variation. We also showed that the more criteria chosen, the worse the quality scores appeared to be. This could be a reflection of the difficulty of complex consultations and the need to prioritise tasks or an unintended consequence of the coders' ability to choose the number of criteria to apply. It seems clear that this issue should be mitigated in future studies by two possible strategies. Focusing on a single presenting symptom would allow a single set of criteria to be used. Alternatively, we would recommend instructing coders to assign a patient specific set of criteria relevant to the presenting symptom(s) as determined by study staff or a clinical supervisor, instead of leaving the choice of criterion relevant to the consultation up to the raters.

Low reliability can result from large amounts of noise in the measurement or small differences between the targets of measurement or both [60]. Thus, a second and less appreciated reason for low reliability of quality measurement is if the observed population does not vary much in the quality of care received. For example, if encounters at outpatient academic hospitals are of consistently high quality, the reliability of a quality of care measurement procedure estimated only in that population will be lower than in a population receiving more heterogeneous levels of quality of care. We might well find that the reliability is much better when trying to distinguish quality of care received across the more varied sites of care representing the breadth of facilities where a target patient population actually receives care. Reliability is most relevant when estimated in the same population for which the measurement is intended. Thus, future studies should define the target population of people and care settings carefully.

The third important issue is the rater population. Using this measurement procedure, the true score of quality of care that we identify is the average score that would be estimated using an infinite number of raters selected from the same population from which we selected our raters. In a larger study or operational system, it would be important to carefully define the population of raters to which we wanted our scores to generalise. In our study, discrepancies could have emerged between coders relating to their differing levels of experience: one was a retired expert physician, while the other was a more junior physician. Defining the target population of raters (e.g. experts vs. senior physicians vs. community providers) is a normative decision for which there is still a lack of clarity or guidance in the existing literature [22, 55]. Our findings illustrate the need to further examine and resolve this issue.

### Strengths and limitations

To our knowledge, this is the first study to use video-observation to examine provider quality in a LMICs hospital outpatient setting, which is a particular strength of this work. Further consideration should be made to the measurement of provider quality in this setting given the dearth of literature that currently exists. A limitation of our study is the modest sample size. However, this reflects our intent to do a preliminary examination of feasibility and acceptability of video-observation for assessing technical quality of care and not to provide precise estimates of encounter quality at the patient, provider, or encounter level. The estimates of provider quality we obtained were considerably higher than those reported in prior studies in community or clinic-based primary healthcare settings in LMICs [27, 30, 31, 61–64]. This could either reflect high quality of care in the University-based practice setting we studied or be due to overly generous raters. Raters independent of the hospital and blinded to the setting will ensure that there is no “home-team” bias in scoring. In addition, we would expect more heterogeneity in scores when more diverse settings and provider types are studied and in multi-site studies. A limitation of video-observation is that similar to all forms of direct observation, video is inherently cross-sectional and cannot evaluate processes of care for an entire episode of illness, from presentation to resolution of symptoms or death. Case-note review or outcome-based measurement are methods that attempt to capture the care for an entire episode, but each have their own measurement challenges [58, 65, 66] and are particularly difficult in LMICs, where case-notes often lack sufficient detail and outcome data is not routinely collected.

### Implications and conclusion

Our study shows that video-observation is feasible and acceptable to implement in a hospital outpatients’ department in an LMICs setting, to examine the technical quality of primary care provision. However, further examinations of the acceptability of this method from the perspective of patients and providers across a broad array of settings are required. We also found that there are caveats in the use of video-observation and necessary improvements to be made to our approach. Although our study was small and preliminary, it does raise the possibility that a rating based on expert global judgement might be sufficient to assess video encounters for aggregated quality assessments at the clinic and community level and more flexible and simpler to maintain than an explicit process score. It also seems likely that low-income settings may be distinguished not just by lower overall quality but more variability in quality than high income settings. Larger scale studies across sites of care that are representative of the heterogeneity of quality of care found in a community or country are required to assess whether the procedure has sufficient reliability to practically monitor quality at the clinic or hospital level.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12913-021-06491-4>.

#### Additional file 1.

### Acknowledgements

We would like to thank the physicians and patients at UCH in Ibadan, Nigeria who took part in this study. We thank colleagues at the University of Ibadan and UCH for facilitating this research. We extend our thanks and appreciation to Dr Tawakalit Olubukola Salaam (Consultant Family Physician), who led the process of identifying eligible patients in the clinic. Upon acceptance, NA had moved to Newcastle University, Population Health Sciences Institute and may be contacted on [Nav.Aujla@newcastle.ac.uk](mailto:Nav.Aujla@newcastle.ac.uk).

### Authors’ contributions

All authors helped conceive and design the study. MMA and SOB recruited participants and collected the data. TI, AI, AO1 (University College Hospital (UCH), Ibadan) and EO, OF, AO2 (University of Ibadan) managed and supervised the fieldwork. RJJ and NA (University of Warwick) supported from the UK. SIW conducted the statistical analyses. All authors helped interpret the data, with specialist input on the statistical results from TPH. NA wrote the first draft of the manuscript and worked with co-authors to prepare subsequent versions. TPH and RJJ were major contributors to writing subsequent versions of the manuscript. All authors reviewed the manuscript and approved the final version.

### Funding

This research was funded by two grants awarded by the University of Warwick, UK (Global Research Priorities – International Development (GRP - ID) and GCRF Accelerator Fund). The research is part of the National Institute for Health Research (NIHR) Global Health Research Unit on Improving Health in Slums funded using UK aid from the UK Government to support global health research. The authors, NA, EO, OF, MMA, SIW, AO2 and RJJ were supported by this NIHR grant. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the UK Department of Health and Social Care. The funder had no role in the design

of the study and collection, analysis, and interpretation of data and in writing the manuscript.

#### Availability of data and materials

Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available for this study.

#### Declarations

##### Ethics approval and consent to participate

Ethical approval was granted by the University of Warwick Biomedical and Scientific Research Ethics Committee (REGO-2018-2306) and the University of Ibadan/UCH, Ibadan Research Ethics Committee (UI/EC/18/0646). Adult patients and parents/caregivers of child patients provided written informed consent to take part in this study.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>Warwick Medical School, University of Warwick, C/O Room B147a, CV4 7AL Coventry, United Kingdom. <sup>2</sup>University College Hospital, Ibadan, Nigeria. <sup>3</sup>University of Ibadan, Ibadan, Nigeria. <sup>4</sup>Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom. <sup>5</sup>Department of Medicine, University of Michigan Medical School, Ann Arbor, MI, USA.

Received: 12 October 2020 Accepted: 6 May 2021

Published online: 22 May 2021

#### References

- World Health Organization. Declaration of Astana. Global Conference on Primary Health Care, Astana, Kazakhstan 25–26 October 2018. World Health Organization and the United Nations Children's Fund (UNICEF); 2018.
- World Health Organisation. Declaration of Alma-Ata. International Conference on Primary Health Care, Alma-Ata, USSR 6–12 September 1978. Geneva: WHO; 1978.
- Chabot J. The Bamako Initiative. *Lancet*. 1988;10(2(8624)):1366–7.
- World Health Organization. Thirteenth General Programme of Work. Promote health, keep the world safe, serve the vulnerable. Geneva: World Health Organization; 2019.
- United Nations. Transforming our world: the 2030 agenda for Sustainable Development. United Nations; 2015.
- Kruk ME, Gage AD, Arsenault C, Jordan K, Leslie HH, Roder-DeWan S, et al. High-quality health systems in the Sustainable Development Goals era: time for a revolution. *Lancet Glob Health*. 2018;6(11):e1196–e252.
- National Academies of Sciences Engineering and Medicine. Crossing the global quality chasm: improving health care worldwide. Washington, DC: The National Academies Press; 2018.
- World Health Organization. Organisation for Economic Co-Operation and Development, and The World Bank. Delivering quality health services: a global imperative for universal health coverage. 2018. Licence: CC BY-NC-SA 3.0 IGO.
- Hrisos S, Eccles MP, Francis JJ, Dickinson HO, Kaner EFS, Beyer F, et al. Are there valid proxy measures of clinical behaviour? A systematic review. *Implement Sci*. 2009;4(1):37.
- Donabedian A. Evaluating the quality of medical care. *The Milbank Memorial Fund Quarterly*. 1966;44(3):166–203.
- Lilford RJ, Chilton PJ, Hemming K, Girling AJ, Taylor CA, Barach P. Evaluating policy and service interventions: framework to guide selection and interpretation of study end points. *BMJ*. 2010;341:c4413.
- Donabedian A. The quality of care: how can it be assessed? *JAMA*. 1988; 260(12):1743–8.
- Lilford RJ, Brown CA, Nicholl J. Use of process measures to monitor the quality of clinical practice. *BMJ*. 2007;335(7621):648–50.
- Brown CA, Hofer T, Johal A, Thomson R, Nicholl J, Franklin BD, et al. An epistemology of patient safety research: a framework for study design and interpretation. Part 3. End points and measurement. *Qual Saf Health Care*. 2008;17(3):170–7.
- Luna D, Otero C, Marcelo A. Health informatics in developing countries: systematic review of reviews. Contribution of the IMIA Working Group Health Informatics for Development. *Yearb Med Inform*. 2013;8:28–33.
- Rowe AK, Onikpo F, Lama M, Deming MS. Evaluating health worker performance in Benin using the simulated client method with real children. *Impl Sci*. 2012;7(1):95.
- Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;65(9):63–7.
- Sommer R. The Hawthorne Dogma. *Psychol Bull*. 1968;70 (6 (Pt.1)):592–5.
- Leonard K, Masatu MC. Outpatient process quality evaluation and the Hawthorne Effect. *Soc Sci Med*. 2006;63(9):2330–40.
- Cardemil CV, Gilroy KE, Callaghan-Koru JA, Nsona H, Bryce J. Comparison of methods for assessing quality of care for community case management of sick children: an application with community health workers in Malawi. *Am J Trop Med Hyg*. 2012;87(5 Suppl):127–36.
- Franco LM, Franco C, Kumwenda N, Nkhoma W. Methods for assessing quality of provider performance in developing countries. *Int J Qual Health Care*. 2002;14(Suppl 1):17–24.
- Hermida J, Nicholas DD, Blumenfeld SN. Comparative validity of three methods for assessment of the quality of primary health care. *Int J Qual Health Care*. 1999;11(5):429–33.
- Onishi J, Gupta S, Peters DH. Comparative analysis of exit interviews and direct clinical observations in pediatric ambulatory care services in Afghanistan. *Int J Qual Health Care*. 2011;23(1):76–82.
- Miller NP, Amouzou A, Hazel E, Degefe T, Legesse H, Tafesse M, et al. Assessing the quality of sick child care provided by community health workers. *PLoS ONE*. 2015;10(11):e0142010.
- Das J, Hammer J. Money for nothing: the dire straits of medical practice in Delhi, India. *J Dev Econ*. 2007;83(1):1–36.
- Das J, Hammer J. Quality of primary care in low-income countries: facts and economics. *Annu Rev Econ*. 2014;6(1):525–53.
- Das J, Hammer J, Leonard K. The quality of medical advice in low-income countries. *J Econ Perspect*. 2008;22(2):93–114.
- Das J, Holla A, Mohpal A, Muralidharan K. Quality and accountability in health care delivery: audit-study evidence from primary care in India. *Am Econ Rev*. 2016;106(12):3765–99.
- Das J, Mohpal A. Socioeconomic status and quality of care in rural India: new evidence from provider and household surveys. *Health Aff*. 2016;35(10):1764–73.
- Kwan A, Daniels B, Saria V, Satyanarayana S, Subbaraman R, McDowell A, et al. Variations in the quality of tuberculosis care in urban India: a cross-sectional, standardized patient study in two cities. *PLoS Med*. 2018;15(9): e1002653.
- Daniels B, Dolinger A, Bedoya G, Rogo K, Goicoechea A, Coarasa J, et al. Use of standardised patients to assess quality of healthcare in Nairobi, Kenya: a pilot, cross-sectional study with international comparisons. *BMJ Global Health*. 2017;2(2).
- Eke CB, Ibekwe RC, Muoneke VU, Chinawa JM, Ibekwe MU, Ukoha OM, et al. End-users' perception of quality of care of children attending children's outpatients clinics of University of Nigeria Teaching Hospital Ituku-Ozalla Enugu. *BMC Res Notes*. 2014;7:800.
- Nabbuye-Sekandi J, Makumbi FE, Kasangaki A, Kizza IB, Tugumisirize J, Nshimye E, et al. Patient satisfaction with services in outpatient clinics at Mulago Hospital, Uganda. *Int J Qual Health Care*. 2011;23(5):516–23.
- Ajayi IO, Olumide EA, Oyediran O. Patient satisfaction with the services provided at a general outpatients' clinic, Ibadan, Oyo State, Nigeria. *Afr J Med Med Sci*. 2005;34(2):133–40.
- Kabatooro A, Ndoboli F, Namatovu J. Patient satisfaction with medical consultations among adults attending Mulago Hospital Assessment Centre. *South Afr Fam Pract*. 2016;58(3):87–93.
- Puri N, Gupta A, Aggarwal AK, Kaushal V. Outpatient satisfaction and quality of health care in North Indian Medical Institute. *Int J Health Care Qual Assur*. 2012;25(8):682–97.
- Juma D, Manongi R. Users' perceptions of outpatient quality of care in Kilosa District Hospital in Central Tanzania. *Tanzan J Health Res*. 2009;11(4): 196–204.
- Byrne PS, Long BEL. Doctors talking to patients. Exeter: RCGP; 1976.
- Zimmer KP, Solomon BS, Siberry GK, Serwint JR. Continuity-structured clinical observations: assessing the multiple-observer evaluation in a pediatric resident continuity clinic. *Pediatrics*. 2008;121(6):e1633–45.

40. Federal Ministry of Health. Standard treatment guidelines (Nigeria). Nigeria: Federal Ministry of Health; 2008.
41. World Health Organization. Integrated Management of Childhood Illness: A WHO/UNICEF Initiative. WHO Bulletin: World Health Organization; 1997.
42. World Health Organization. IMAI District Clinician Manual: hospital care for adolescents and adults (Volume 2): Guidelines for the management of illnesses with limited-resources. Geneva: WHO; 2011.
43. Rubenstein LV, Kahn KL, Harrison ER, Sherwood MJ, Rogers WH, Brook RH. A RAND Note. Structured implicit review of the medical record: a method for measuring the quality of in-hospital medical care and a summary of the quality changes following implementation of the Medicare prospective payment system. Santa Monica: RAND; 1991.
44. Rubenstein LV, Kahn KL, Reinisch EJ, Sherwood MJ, Rogers WH, Kamberg C, et al. Changes in quality of care for five diseases measured by implicit review, 1981 to 1986. *JAMA*. 1990;264(15):1974–9.
45. Keeler EB, Rubenstein LV, Kahn KL, Draper D, Harrison ER, McGinty MJ, et al. Hospital characteristics and quality of care. *JAMA*. 1992;268(13):1709–14.
46. Kerr EA, Hofer TP, Hayward RA, Adams JL, Hogan MM, McGlynn EA, et al. Quality by any other name?: A comparison of three profiling systems for assessing health care quality. *Health Serv Res*. 2007;42(5):2070–87.
47. Hutchinson A, Coster JE, Cooper KL, McIntosh A, Walters SJ, Bath PA, et al. Assessing quality of care from hospital case notes: comparison of reliability of two methods. *Qual Saf Health Care*. 2010;19(6):e2-e.
48. Ashton CM, Kuykendall DH, Johnson ML, Wray NP. An empirical assessment of the validity of explicit and implicit process-of-care criteria for quality assessment. *Med Care*. 1999;37(8):798–808.
49. Goldstein H, Browne W, Rasbash J. Partitioning variation in multilevel models. *Underst Stat*. 2002;1(4):223–31.
50. Crabtree B, Miller W. Using codes and code manuals: a template organizing style of interpretation. In: Crabtree B, Miller W, editors. *Doing qualitative research*. 2nd ed. Thousand Oaks: Sage Publications; 1999. pp. 163–77.
51. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77–101.
52. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Measur*. 1973;33(3):613–9.
53. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*. 1977;33(2):363–74.
54. Kahn KL, Rubenstein LV, Draper D, Kosecoff J, Rogers WH, Keeler EB, et al. The effects of the DRG-based prospective payment system on quality of care for hospitalized Medicare patients: an introduction to the series. *JAMA*. 1990;264(15):1953–5.
55. McCarthy KJ, Blanc AK, Warren CE, Mdawida B. Women's recall of maternal and newborn interventions received in the postnatal period: a validity study in Kenya and Swaziland. *J Glob Health*. 2018;8(1):010605.
56. Rethans J-J, Gorter S, Bokken L, Morrison L. Unannounced standardised patients in real practice: a systematic literature review. *Med Educ*. 2007;41(6):537–49.
57. Aujla N, Chen Y-F, Samarakoon Y, Wilson A, Grolmusová N, Ayorinde A, et al. Comparing the use of direct observation, standardised patients and exit interviews in low-and middle-income countries: a systematic review of methods of assessing quality of primary care. *Health Policy Plan*. 2020: czaa152.
58. Hofer TP, Asch SM, Hayward RA, Rubenstein LV, Hogan MM, Adams J, et al. Profiling quality of care: Is there a role for peer review? *BMC Health Serv Res*. 2004;4(1):9.
59. Spearman C. Correlation calculated from faulty data. *Br J Psychol*. 1910;3(3):271–95.
60. Dunn G. *Statistical evaluation of measurement errors: design and analysis of reliability studies*. 2nd edition. London: Arnold; New York: Distributed in the United States of America by Oxford University Press; 2004.
61. Das J, Holla A, Das V, Mohanan M, Tabak D, Chan B. In urban and rural India, a standardized patient study showed low levels of provider training and huge quality gaps. *Health Aff*. 2012;31(12):2774–84.
62. Das J, Kwan A, Daniels B, Satyanarayana S, Subbaraman R, Bergkvist S, et al. Use of standardised patients to assess quality of tuberculosis care: a pilot, cross-sectional study. *Lancet Infect Dis*. 2015;15(11):1305–13.
63. Das J, Sohnesen TP. Variations in doctor effort: evidence from Paraguay. *Health Aff*. 2007;26(3):w324–37.
64. Daniels B, Kwan A, Pai M, Das J. Lessons on the quality of tuberculosis diagnosis from standardized patients in China, India, Kenya, and South Africa. *J Clin Tuberc Other Mycobact Dis*. 2019;16:100109.
65. Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. *Statist Sci*. 2007;22(2):206–26.
66. Lilford R, Edwards A, Girling A, Hofer T, Di Tanna GL, Petty J, et al. Inter-rater reliability of case-note audit: a systematic review. *J Health Serv Res Policy*. 2007;12(3):173–80.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

