Research Articles

Implementing electronic data collection platform for household surveys in resource-constrained settings

Prasanna Samuel Premkumar[1], Santhosh Kumar Ganesan[1], Balaji Pandiyan[1], Dhivya Kumari Krishnamoorthy[1]

[1] The Wellcome Trust Research Laboratory, Division of Gastrointestinal Sciences, Christian Medical College, Vellore, Tamil Nadu, India

Keywords: health information systems, mobile technologies, open data kit, household surveys

https://doi.org/10.29392/001c.19053

Journal of Global Health Reports

Vol. 5, 2021

Background

In resource-constrained settings, quality and timeliness of data are the main concerns related to the use of information systems for decision making. Many different tools are available to improve such systems, but their usefulness is only been recently explored. In this paper, we describe our implementation of an electronic platform, open data kit (ODK) for data collection and its feasibility in data management for a population-based household health expenditure survey.

Methods

We evaluated the use of ODK based data collection in households located in two areas (one urban and one rural) in Vellore, Tamil Nadu, India. From each area, we selected a sample of 60 households for piloting the ODK based questionnaires. The household survey questionnaires were programmed using the Microsoft Excel for data collection in the ODK collect android application. The ODK aggregate was used for data storage and data transfer. A team of six field workers was recruited, and trained to use the ODK collect application for survey data collection. After the training, the field workers pilot tested the questionnaires, both in the form of mock surveys and real on-field testing.

Results

Under mock-interviews, there were no significant differences in time –to completion between the six field workers. A total of 60 households participated in field testing that showed field workers were able to complete the questionnaires in a timely manner, (mean 32 minutes (SD=18)) with minimal errors, and all field workers found the ODK form easy to use. There were no major technical issues in the ODK implementation or with electronic devices.

Conclusions

Results from both mock interviews and on-field testing of our data collection platform show the feasibility of using this approach in resource-constrained settings. The approach used to implement, integrate, and test this platform can benefit other health researchers in developing settings intending to move from paper-based methods toward electronic data collection systems.

High quality and timely data are major concerns for organizations in many resource-constrained settings in making better decisions.[1] Data collection modes play an integral role in influencing both the timeliness and quality of data.[2] Traditionally, paper-based methods for data collection are preferred but require a great amount of person-hours to be spent on data-collection, entry and cleaning. This approach is time-consuming, error prone, and challenging to accurately monitor field operations, participant refusals, and perform quality control. Advancements in information and communication technology (ICT) have led to development of mobile-based technology for data collection to overcome some of these inherent limitations with paper-based data collection.[3–5] With mobile-based data collection, data is captured with adequate validation at the level of data entry and transferred digitally reducing the chance of human error leading to more accurate and timely data.

Many public health and development research groups are exploring various strategies to develop applications for data collection in health and development research.[6] Most of these applications have an in-built point of source data consistency checks, linked with servers and dynamic web interfaces (dashboards), which can present a visual summaries in real-time. Various mobile applications (e.g. Epi-Collect, EpiSurveyor, KoBo, RapidSMS, Open Data Kit (ODK), CSPro, etc.,) for data collection in surveillance, clinical research, and survey are available,[7] however their usefulness is only recently explored in resource-constrained settings. While numerous studies agree on improvement of efficiency through mobile based systems,[8,9] the implementation of these systems is often challenged by common barriers such as lack of adequate prior knowledge and experience, perceived reliance on programming expertise, and defunct user-interface interaction. However, to date, few studies have provided information useful for implementing the mobilebased tools for data collection in large population based studies.[10–12] Further as mobile-based tools require interviewers to manage interactions both with the mobile and survey respondent, information on user-interaction of this approach is crucial.

To address these gaps, we evaluated whether ODK based platform will be feasible to implement for data collection in household surveys. In this paper, we aim to (1) describe our implementation of ODK technology for mobile-based data collection in household surveys and (2) report its feasibility for use with survey field workers with little knowledge or prior experience in mobile devices for data collection in household surveys.

METHODS

SURVEY SETTING

This pilot study was part of large survey evaluating household healthcare expenditures. We obtained informed consent and interviewed a small sample of households to gain insights into the use of mobile applications for data collection. We evaluated use of ODK based data collection in households located in two areas (one urban and one rural) in Vellore, Tamil Nadu, India. These areas were among the targets for the main household surveys. From each area, we selected a sample of 60 households for piloting the ODK based questionnaire. The pilot period which included training and testing of field workers was scheduled during May-June 2019.

FEASIBILITY TESTING

We conducted two rounds of formal testing of ODK based data collection. Six field workers and two-supervisors, one data manager, participated in both rounds of testing. The first round consisted of one-day training on the use of digital questionnaires using ODK technology for data collection. It consisted of a two-hour demonstration of the questionnaire in the ODK application, followed by one-hour individual practice. After the training session, fieldworkers were asked to perform sample tasks of obtaining consent, completing household roster, and more advanced features such as the use of branching and skip logic for modules of the utilization of healthcare outpatient visits and hospitalizations using the ODK collect application. Each field worker was assigned these sets of tasks, following which we observed their completion and asked for their impressions on ease in performance of the task.

Following the training session, we conducted an experiment in which four volunteers acted as prospective household survey respondents. The six field workers responded to hypothetical scenarios representing four main branching processes present within the survey that could result in health expenditures: (i) a household with no healthcare utilization or hospitalization, but had general health expenditure, (ii) a household with one or more members having recent illness and utilized day healthcare services only, (iii) a household with one or members recently hospitalized only, and (iv) a household with one or more members both utilized day healthcare and hospitalized and (**Figure 1**). All six field workers conducted interviews on the four volunteers and recorded their responses in the tablets. We hypothesized that no differences in the time to completion between the six field workers for each scenario, given the baseline training sessions.

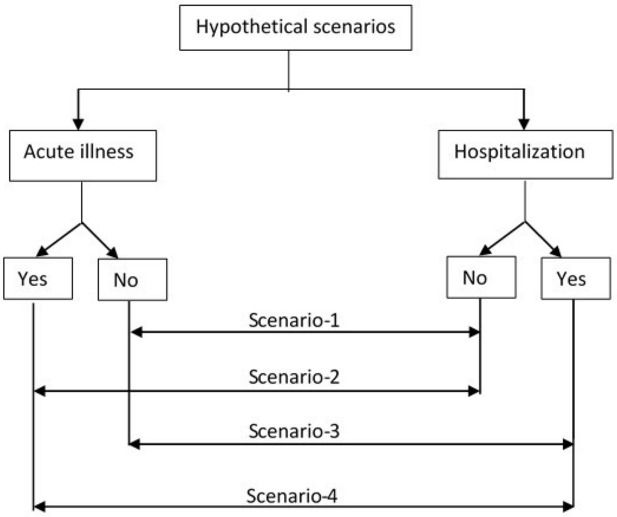

Figure 1. Diagram representing four hypothetical scenarios for data collection.

In the second round, each field worker was asked to recruit a target sample of five households. The data collection was performed under supervision, timed and the magnitude of difference between households with sub-categories/characteristics was noted to arrive at an expected time to complete the survey. After completion of each household survey, field workers uploaded the form into the ODK aggregate server.

SURVEY INSTRUMENT

The original questionnaire composed a total of 142 questions to assess households total and health expenditures which were selected from national sample surveys on household expenditures on various items under different categories.[13] Respondents are asked to report how much they spent on different categories of goods and services within a certain period. Questions were grouped by the following sections: (i) household sampling and identification – 18 questions, (ii) roster – 22 questions, (iii) illness history, recent outpatient visits, and health expenditure – 32 questions, (iv) hospitalization and related expenditure – 32 questions, (v) socio-economic status – 12 questions and (vi) household expenditure – 36 questions. And response classes were either being single/multiple choices, integer, date or open-ended. Filters were used to prevent a respon-

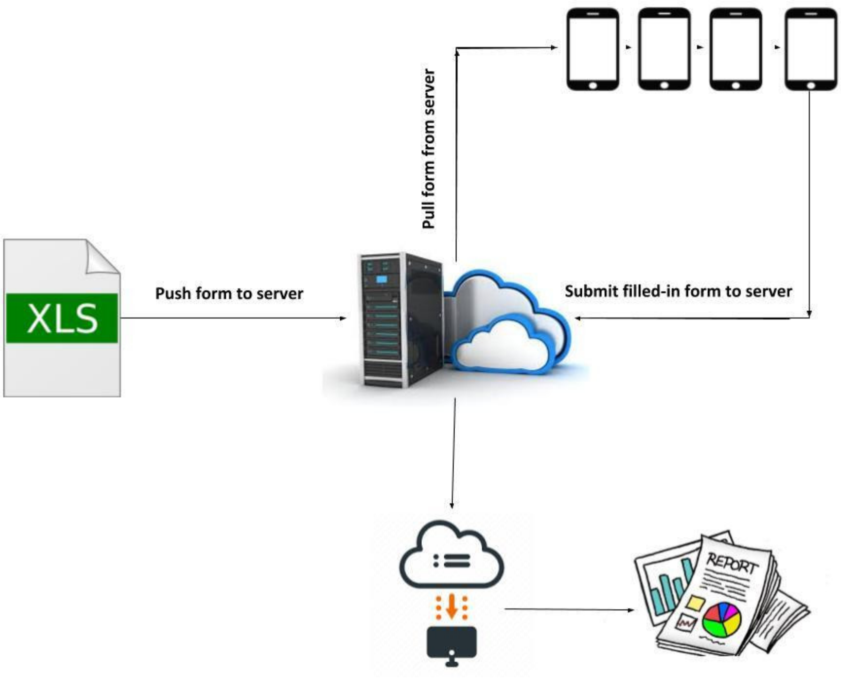

Figure 2. Workflow in mobile based data collection.

dent from being asked questions from sections that are irrelevant or not applicable to the individual households.

For household identification, we assigned alpha-numeric text which conveyed phase and sites at which households are enrolled into the study. The first digit conveyed the phase of data collection and the subsequent five digits conveyed spatial location details of the households: (i) urban or rural, (ii) clusters, (ii) fieldworker in-charge, and (iv) a two-digit serial no for the household. To link individual members within the household, we added two-digits to the household ID number. For example, FUC08110 indicates data representing first phase, urban region, cluster 08, fieldworker 1, and household number 10.

TOOLSET FOR DIGITAL DATA COLLECTION

For the design of data collection forms, we evaluated ODK tools, notably ODK build for form designer, ODK collect for user-interface, and ODK aggregate for storage services. We used MS Excel for the design of mobile forms, converted to Xforms format using XLSFORM (https://xlsform.org) online tool, and validated using Enketo (https://enketo.org/xforms).

The validated form was uploaded onto the ODK aggregate server, a web-based application to support mobile data collection and provides hosting services for data collected through mobile devices (**Figure 2**). The coded form is available online here (https://github.com/HES-Project/Documents/blob/main/HES_form.xls).

The field workers used a tablet device (Samsung Galaxy Tab A (T285N)) installed with ODK collect app., and downloaded the digital forms for data collection from the ODK aggregate server onto their devices for the survey. After surveying the households with tablets, the completed forms were submitted back to the ODK aggregate server. The data manager, tracked submissions online, reviewed submissions in real-time or near real-time and then downloaded the data in comma-separated values (CSV) format for further analysis.

OUTCOMES

The main outcomes for the study are completion and error rates, time-to-completion, and reported user comments.

DATA ANALYSIS

Average time to completion of surveys was calculated during both mock interviews and pilot survey. In pilot surveys, we calculated number of responses that had either erroneous, inconsistent responses or missing entries. All analyses were performed using STATA V12 (Statacorp, Texas, US).

RESULTS

FIELD WORKERS TRAINING AND MOCK INTERVIEWS

Six field workers along with two supervisors participated in a one-day training session. Followed by the training, field workers and supervisors undertook mock interview sessions. In the mock sessions, each fieldworker administered the questionnaire to four volunteers who represented the four above mentioned hypothetical scenarios. Scenario 1, 2, and 3 took on average about 27, 27 and 25 minutes (standard deviation, SD=2, SD=2, SD=3), respectively. Scenario 4 in which both out-patient and in-patient visits were present took the maximum time for completion (mean=45 minutes, SD=4). And all field workers reported that they were able to use the application effectively, and were satisfied with the

Table 1. Summary statistics of survey time in different scenarios.

Scenario	Outpatient department	Inpatient department	Number of households	Mean (SD) time taken
1	Absent	Absent	19	24.4 (7.1)
2	Present	Absent	30	31.9 (15.4)
3	Absent	Present	8	43.0 (29.9)
4	Present	Present	3	59.7 (5.5)
All			60	32(18)

SD – standard deviation

design structure of the questionnaire in the ODK application.

All completed forms seamlessly transferred to the ODK aggregate server, and were later exported as CSV files. No major technical issues were observed in downloading data forms from the ODK server to the mobile devices and submitting filled-in forms back to the ODK server. There were no hardware or software failures using the tablets.

PILOT FIELD SURVEY

We enrolled 60 households, 30 from urban and 30 from rural localities. All households consented to participate in the survey. All entries had an accurate date and time recordings for when the data collection occurred. Of these 60 samples, errors or inconsistencies of responses were observed in recording Global Positioning System (GPS) locations of households, dates of outpatient visits and inpatient admissions, and in few categories of household expenditure which had very large entries. Few open-ended responses on reasons for outpatient visits or inpatient admissions also resulted in inconsistent or erroneous responses.

Similarly, missing responses were observed in the following items: date of birth, GPS locations, contact numbers for alternative persons, and monthly income. There was also general agreement, regardless of locality, that questionnaire items on large expenditures, including details on savings and taxation were the most problematic part of the survey questionnaire. Table 1 shows the average times taken to survey completion for households with/without out-patient and hospitalizations. The average duration of data collection for the pilot household survey was 32 minutes (SD=18 minutes). Quality checks were performed in real-time, and inconsistencies were detected, rectified, and cleaned in a timely manner. The real-time submission of completed surveys provided ongoing data on interview start time, end time, and time taken to complete each survey. The automated graphs and reports from the ODK web-interface allowed the data manager to visualize outputs such as survey completion count on an hourly or daily basis or the average survey completion time.

DISCUSSION

In this pilot study, we used ODK, an android application to enter and upload data at the point of collection from participants of the household survey. The android application, ODK collect paired with the web-based data management platform ODK aggregate enabled real-time supervision of field workers and helped to reduce error rates.

The major advantage of using the ODK system was that data cleaning efforts post field work were minimal. This was mainly due to the inclusion of validation and consistency checks in the digitized form, which prevented erroneous data being entered into the form at the point of collection. This meant not only an improvement in the overall accuracy of the data, but a huge time saving in the time previously spent for processing datasets, identifying inconsistencies, and then undertaking corrective measures.

The other advantage was that it enhanced supervisory activities to operational guidelines that were to be followed by the field workers. For example, field workers reported that meta-data from activity logs provided a real-time summary of field activities and made them keener on ensuring they are making field visits as required. Further, automated time-stamps and location data collected at the time of entry played an important role in increasing the validity of collected data. Although we did not use paper-based methods for comparisons, it is well known from previous studies that turn-around time for paper-based methods are much longer and requires different levels of supervisory assessments to reduce errors and ensure high-quality data.[14–16]

CHALLENGES (OR UNCERTAINTY) WITH DIGITAL DATA COLLECTION AND LESSONS LEARNT

Major challenges were related with GPS for capturing locations, text-based responses, navigation back and forth through the questionnaire. We noted problems with recording the location of households when the GPS signals were weak. When interviewers proceeded with the survey without confirming the proper location of households, location information was marked grossly wrong. To avoid this in the main study, a protocol was developed to calibrate devices and increase devise responsiveness before recording the location information. Additionally, there was one programming error in which skip patterns were set incorrectly that resulting in invalid responses.

The other problems observed in the forms were mainly present in open-ended questions which required textual responses. For textual responses, we could not restrict or set a limit for the entries. This resulted in invalid or erroneous responses. Further, fieldworkers reported difficulty in typing out the response for these questions, while other types of questions required just a few checkmarks or entering numbers in the device. In the main study, we overcome this problem by reducing the total number of questions requiring textual responses to a minimal level. Missing responses were largely present in questions from expenditure sections. During the interviews, participants did not want to respond to questions related to their taxation, savings, and borrowings. This is expected owing to the personal nature of these questions and might have been more demanding to the household respondent.[17,18]

The main limitation was the length of the questionnaire and the longer time required from participants for completion of the survey. Due to the length of the questionnaire, the household interviews were performed only once. Ideally, the questionnaires could be administered by more than one interviewer to allow for more comparisons regarding the efficiency of field workers. Though ODK remained stable and responsive, the length affected interviewers to navigate back and forth through the different sections, subsections, and items in the forms. The forms were set with a one-question prompt at a time and some of the fields were dynamic such that responses from prior questions were used in subsequent questions, so if the interviewer had to go back to sections, then it resulted in delays and caused inconvenience to the household participants. We attempted to manage this limitation by our in-house training and mock interviews for field workers to limit such situations.

CONCLUSIONS

Electronic data collection is being increasingly used for data collection, though not without its challenges and problems. We believe this study will help other researchers from developing settings in effectively integrating open source tools for electronic data collection and management within their research studies.

ACKNOWLEDGMENTS

We would like to thank field supervisors (Mr Arumugam and Mr Sathish) and all field workers (Ms Saranya, Ms Kannagi, Ms Muthulakshmi, Mr Santhosh, Mr Sivaprasad and Mr Jagadeesh) involved with the conduct of the study. We are grateful to Professor Gangandeep Kang for her support and Dr Venkata Ragaha Mohan for initial insight on use of open data solutions for data collection. We thank the survey participants for their cooperation.

FUNDING

The household survey was supported by Science and Engineering Research Board (SERB) (Grant No: EMR/2017/001864), Department of Science & Technology (DST), New Delhi.

AUTHOR CONTRIBUTIONS

PSP conceived the study, wrote and revised the manuscript. SG, BP, DK assisted with data collection, analysis and input to writing the manuscript and provided revisions to it. All authors read and approved the final manuscript.

COMPETING INTERESTS

The authors completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf, and declare no conflicts of interest.

CORRESPONDENCE TO:

Dr. Prasanna Samuel Premkumar, PhD
The Wellcome Trust Research Laboratory
Division of Gastrointestinal Sciences
Christian Medical College, Vellore,
Tamil Nadu 632004, India.
prasanna.samuel@cmcvellore.ac.in

Submitted: November 30, 2020 GMT, Accepted: January 30, 2021 GMT

REFERENCES

1. Boulos MNK, Wheeler S, Tavares C, Jones R. How smartphones are changing the face of mobile and participatory healthcare: An overview, with example from eCAALYX. *Biomed Eng Online*. 2011;10(1):24. doi:10.1186/1475-925x-10-24

2. Anokwa Y, Hartung C, Brunette W, Borriello G, Lerer A. Open source data collection in the developing world. *Computer*. 2009;42(10):97-99. doi:10.1109/mc.2009.328

3. Aanensen DM, Huntley DM, Feil EJ, al-Own F, Spratt BG. EpiCollect: Linking smartphones to web applications for epidemiology, ecology and community data collection. Hay SI, ed. *PLoS ONE*. 2009;4(9):e6968. doi:10.1371/journal.pone.0006968

4. Guo TW, Laksanasopin T, Sridhara AA, Nayak S, Sia SK. Mobile device for disease diagnosis and data tracking in resource-limited settings. *Methods Mol Biol Clifton NJ*. 2015;1256:3-14. doi:10.1007/978-1-4939-2172-0_1

5. Haskew J, Kenyi V, William J, et al. Use of mobile information technology during planning, implementation and evaluation of a polio campaign in South Sudan. Pett SL, ed. *PLoS ONE*. 2015;10(8):e0135362. doi:10.1371/journal.pone.0135362

6. Hartung C, Lerer A, Anokwa Y, et al. Open data kit: Tools to build information services for developing regions. In: *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*. ; 2010:1-12.

7. King C, Hall J, Banda M, et al. Electronic data capture in a rural African setting: Evaluating experiences with different systems in Malawi. *Glob Health Action*. 2014;7(1):25878. doi:10.3402/gha.v7.25878

8. Ganesan M, Prashant S, Jhunjhunwala A. A review on challenges in implementing mobile phone based data collection in developing countries. *J Health Inform Dev Ctries*. 2012;6(1).

9. King JD, Buolamwini J, Cromwell EA, et al. A novel electronic data collection system for large-scale surveys of neglected tropical diseases. Noor AM, ed. *PLoS ONE*. 2013;8(9):e74570. doi:10.1371/journal.pone.0074570

10. Singh H. Mobile data collection using an android device. *IJCST*. 2013;4(1):200-202.

11. Wamwenje SAO, Wangwe II, Masila N, Mirieri CK, Wambua L, Kulohoma BW. Community-led data collection using Open Data Kit for surveillance of animal African trypanosomiasis in Shimba hills, Kenya. *BMC Res Notes*. 2019;12(1):151. doi:10.1186/s13104-019-4198-z

12. Kaplan WA. Can the ubiquitous power of mobile phones be used to improve health outcomes in developing countries? *Glob Health*. 2006;2(1):9.

13. National Sample Survey Office. Ministry of Statistics & Programme Implementation, Government of India, New Delhi; 2019.

14. Maduka O, Akpan G, Maleghemi S. Using Android and Open Data Kit Technology in Data Management for Research in Resource-Limited Settings in the Niger Delta Region of Nigeria: Cross-Sectional Household Survey. *JMIR MHealth UHealth*. 2017;5(11):e171. doi:10.2196/mhealth.7827

15. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Med*. 2005;2(10):e267. doi:10.1371/journal.pmed.0020267

16. Tom-Aba D, Olaleye A, Olayinka AT, et al. Innovative technological approach to Ebola virus disease outbreak response in Nigeria using the open data kit and form hub technology. Harper DM, ed. *PLoS ONE*. 2015;10(6):e0131000. doi:10.1371/journal.pone.0131000

17. Tomlinson M, Solomon W, Singh Y, et al. The use of mobile phones as a data collection tool: A report from a household survey in South Africa. *BMC Med Inform Decis Mak*. 2009;9(1):1-8. doi:10.1186/1472-6947-9-51

18. Lori JR, Munro ML, Boyd CJ, Andreatta P. Cell phones to collect pregnancy data from remote areas in Liberia. *J Nurs Scholarsh*. 2012;44(3):294-301. doi:10.1111/j.1547-5069.2012.01451.x