Implementing automated subject selection methodology in an epidemiological survey using open source application, Open Data Kit (ODK)

Ganeshkumar Parasuraman* a, Sabarinathan Ramasamy a, Chokkalingam D a, Vinay Urs b, Ritvik Amarchand c, Sureshkumar N c, Dineshkumar H a, Chandrasekar R a, Naveen Agarwal d

a National Institute of Epidemiology, Chennai, India
b National Centre for Disease Informatics and Research, Bengaluru, India
c All India Institute of Medical Sciences, New Delhi, India
d World Health Organization – SEARO, New Delhi, India

*Corresponding author
Ganeshkumar Parasuraman
ICMR – National Institute of Epidemiology
R127 TNHB, Second Main Road, Ayapakkam,
Chennai 600077, Tamil Nadu, India.
Abstract

Background: Data collection in field-based public health research influences the quality and outcome of the research which in turn aids in developing evidence-based policymaking. As conventional data collection methods are either consume more time or more resources, usage of the cost-effective technology solution is required. Open Data Kit (ODK), the open-source application being designed for an Android-based platform is one such solution to electronic data collection.

Objective: To demonstrate the functionality of ODK in implementing subject selection algorithm Kish methodology in epidemiological research studies.

Methods: Subject selection algorithms in field-based surveys are done manually and one such algorithm is Kish algorithm. Incorporating subject selection algorithm electronically in ODK application was not either tested or reported in the literature. We developed an application framework to build-in Kish grid algorithm in ODK application.

Results: The process and steps for setting up subject selection Kish algorithm to embed in ODK application are explained with syntax function. The application displayed the selected eligible adult on the screen after it runs the syntax function based on the Kish algorithm in the backend.

Conclusion: ODK is an enterprise architecture, which is identified to develop a mobile application for a community based public health survey. The application would reduce the time taken for selecting and completing the interview of study participants. Ever a minor change like spell check on form mandates complete XML generation in ODK application and requires to update the existing entire application.

Keywords: Computing methodologies, Subject Selection, Health Surveys, Data collection, Mobile Applications, Residence characteristics, Open source application, mhealth
Introduction

Collecting, managing and organizing data in field-based epidemiological studies have significant challenges in terms of timeliness of conduct and preparation of analyzable data and quality. [1] Conventional paper-based data collection in surveys has its limitation in both validation and data quality which hinders to generate analyzable electronic data. Also, conventional method would consume significant resources for validation and generating analyzable data. In addition, ensuring data validation during the time of survey would be often impossible in conventional data collection [2]. Thus, in the recent decade, electronic data collection methods gained popularity in field-based surveys. To overcome the challenges of conventional data collection, existing literature supports that electronic data collections methods ameliorate the efficiency of data collection and improve the data quality with a better response rate. [3] In developed countries due to seamless mobile network connectivity and availability of advent usage of electronic devices, real-time data collection and systematization are seamless. This in return will result in high-quality data within the planned time and utilize fewer resources. [4] Generating the research output in time would draw timely evidence-based decisions for public health action.

In developing countries due to lack of capacity, adequate resources and ubiquitous network connectivity, setting up an electronic data management platform for epidemiological surveys are limited. Due to these limitations, most of the organizations practice paper-based data collection methods. [5] Apart from the limitations using an electronic data collection tools in a developing country scenario have other technical hurdles, to name few are expensive licenses, user access restrictions, data security issues. In addition, it may be expensive and complex to develop and deploy on a real-time server under a low capacity setting. Certain open source electronic data collection tools are currently available, and each has different utility. These open source tools may overcome limitations of paper-based data collection tools.

Open Data Kit (ODK) is such a tool which is a free and open-source for mobile data collection is. [6] ODK allows users to design their own form using Microsoft Excel to generate XForms. XForms is an Extensible Markup Language (XML) application that represents the next generation of forms for the Web. [7] In computing, Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. [8] Then the developed XForms is converted into an XML file which required to be uploaded to the ODK server to run the ODK android application on mobile/tablet device. ODK requires little developer knowledge in designing and deploying an android mobile application form for data collection.

Epidemiological surveys use subject selection methodologies from a selected household during field-based data collection. To mention few such study subject selection methodologies are Kish methodology, Age order procedure, Youngest male/Oldest female procedure (YMOF). [9] In a conventional paper-based data collection method selection of study subject by the above methodologies relies totally on manual procedures which may allow errors. Exploiting the electronic tools for study selection methods would definitely abolish manual errors in calculating the selection algorithm.

We designed the ODK collect and utilized ODK aggregate application for National Non-Communicable Disease Monitoring Survey 2017-18. This national survey is being conducted by Indian Council of Medical Research to data on Non-Communicable disease risk factors from 12000 adults across 27 States of India. In this national survey is conducted within 300 urban primary sampling units and 300 rural primary sampling units. The primary sampling unit for a rural area is a
village or group of villages and for an urban area is a census enumeration block or a residential ward. We planned to conduct adult interviews at 20 households per primary sampling unit. We adopted Kish methodology to select one eligible adult per household.

We explored the functionalities of ODK application to incorporate Kish methodology and this paper explains the step by step method with syntax design of ODK application.

Methods

Data management platform framework: ODK Collect renders a survey form, and algorithm into a sequence of input prompts that provide navigation logic, entry constraints, and repeating substructures. Forms are based on the World Wide Web Consortium’s XForms standard and support a wide variety of data types, including GPS coordinates, photos, audio, video, and barcodes. The data entered is stored on the device for asynchronous transfer via the General Packet Radio Service (GPRS), Wi-Fi, or USB cable to any XForms-compatible server.

ODK Aggregate is a ready-to-deploy server that hosts forms and submitted data. [5] It aggregates collected data and provides standard interfaces to extract data in multiple formats such as spreadsheets, queries, and maps. ODK aggregate with other systems and servers via real-time web requests through the internet. We currently implemented ODK aggregate on our own dedicated windows server 2012 – R2 with MySQL database. MySQL database is an open-source relational database management system (RDBMS) released by Oracle Corporation. This also facilitates to host and manage multiple surveys on a single server.

Kish methodology: The Kish selection procedure, developed by the Hungarian born American Statistician Leslie Kish (1910-2000), is considered to be a true random procedure for selecting respondents in the household. The Kish grid or Kish selection grid is a method for selecting members within a household to be interviewed. It uses a pre-assigned table of random numbers to find the person to be interviewed. [10] Understanding of Kish grid algorithm and applying it manually during the survey to select an adult from the household requires intensive training and may prone to errors.

As per the procedure, through a household listing form investigator recorded the name, age and gender of all individuals in a household. As Kish algorithm is used for selecting an adult member of the household, only adult numbers are given a serial number by a specific order. First, the males are numbered in order of decreasing age, followed by the females in the same order. Then the investigator refers a table (Table 1) to select the person to interview. There are six sets of subject selection algorithm based on the number of adults in the household. We prepared a household selection table for the listed 20 households by randomly allocating the Kish selection set to each household (Table 2).
Table 1. Individual selection table of selected household; Kish algorithm

<table>
<thead>
<tr>
<th>Selection set A</th>
<th>Selection set B1</th>
<th>Selection set B2</th>
<th>Selection set C</th>
</tr>
</thead>
<tbody>
<tr>
<td>If # of adults is:</td>
<td>Select adult #</td>
<td>If # of adults is:</td>
<td>Select adult #</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6 or more</td>
<td>6 or more</td>
<td>2</td>
<td>6 or more</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selection set D</th>
<th>Selection set E1</th>
<th>Selection set E2</th>
<th>Selection set F</th>
</tr>
</thead>
<tbody>
<tr>
<td>If # of adults is:</td>
<td>Select adult #</td>
<td>If # of adults is:</td>
<td>Select adult #</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6 or more</td>
<td>6 or more</td>
<td>5</td>
<td>6 or more</td>
</tr>
</tbody>
</table>

Table 2. Household selection table with KISH reference

<table>
<thead>
<tr>
<th>Household</th>
<th>Selection set</th>
<th>Household</th>
<th>Selection set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>11</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>12</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>B1</td>
<td>13</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>B2</td>
<td>14</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>15</td>
<td>B1</td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td>16</td>
<td>B2</td>
</tr>
<tr>
<td>7</td>
<td>D</td>
<td>17</td>
<td>C</td>
</tr>
<tr>
<td>8</td>
<td>D</td>
<td>18</td>
<td>C</td>
</tr>
<tr>
<td>9</td>
<td>E1</td>
<td>19</td>
<td>D</td>
</tr>
<tr>
<td>10</td>
<td>E2</td>
<td>20</td>
<td>D</td>
</tr>
</tbody>
</table>

For example, if the investigator visited household number 1 containing 4 members, then as per table 2, she/he would refer selection set A. As per reference table (Table 1) he would select the first adult listed in the household order by gender and age. This individual will be interviewed further.

Results

Design and incorporation of Kish in ODK: Implementing Kish methodology into ODK application was tricky. After trying out several approaches, we finalized the user-friendly and workable module for Kish in the following steps (Figure 1).

In Step 1, as per Kish methodology, the application allows the user to enter eligible participants with a restriction of a maximum of 6 numbers. In Step 2, the application restricts the user to enter the completed age within the limits set for eligibility criteria of each participant (18 to 69 years). As per
Kish methodology, the application restricts the user to enter the age of the successive household members in the descending order of age. The function syntax used for step 2 is as follows

\[
\text{Function: } ((\text{${gender6}={gender5}\text{ and } .\leq{age5}}) \text{ or } ((\text{${gender6}\neq{gender5}\text{ and } .\geq{minage}\text{ and } .\leq{maxage}})))
\]

In Step 3, Once the user started entering the female household details, the application will not allow male household member details. Thus, Kish algorithm could be successfully implemented in the backend. The constraint function rule for step 3 is as follows

\[
\text{Function: } ((\text{${gender1}=\text{"Female\text{ and } .=\text{"Female\text{ or } \text{(gender1}=\text{"Male\text{)}}}}\text{)}
\]

Where gender1 is previous adult gender and ‘=”Female”’ refers current adult gender. if this condition is not met, then the application will show the following error message “You cannot enter male details after female details.”. Once the adult household details are entered in the application, the application runs the Kish algorithm as explained in Table 1 and 2 in the backend.

As per study protocol, of the 20 households in a primary sampling unit, if visited household number is 7 and the total eligible adults are 5 numbers then application refers to table 2 for selecting the Grid. In this example “set D” will be selected (Table 2).  In our example total eligible adults are 5 numbers thus, the application selects the 4th positioned adults as per “set D” in Table 1.

The syntax for the above example is explained as follows

```java
join(’ ’,${name})        // Concatenate all the loop names into one variable
selected-at($SelName),int(${SelectedRank}-1)) //Select or get selected adult name from an array.
if((${hhnoList}),1,0)    //Syntax for Set A
if((${hhnoList}) and (${tot}<5),1,if((${hhnoList}) and (${tot}>4),2,0)) //Syntax for Set B1
if((${hhnoList}) and (${tot}<4),1,if((${hhnoList}) and (${tot}>3),2,0)) //Syntax for Set B2
if(((${hh_no}=7 or ${hh_no}=8 or ${hh_no}=19 or ${hh_no}=20)) and (${no_available}=1),${no_available},if((((${hh_no}=7 or ${hh_no}=8 or ${hh_no}=19 or ${hh_no}=20) and (${no_available}>1) and (${no_available}<4)),2,if(((${hh_no}=7 or ${hh_no}=8 or ${hh_no}=19 or ${hh_no}=20) and (${no_available}=4)),3,if(((${hh_no}=7 or ${hh_no}=8 or ${hh_no}=19 or ${hh_no}=20) and (${no_available}>4)),4,0)))) // formula for selection in set D
```

After the selection process, the selected adult’s detail such as name, age and sex will be displayed on the screen by the application.
Figure 1: Kish algorithm steps in Open ODK application

Step 1 Household details form screen in the application; Step 2&3 Adult details form screen of the application. Step Application screenshot with the details of the selected adult with age and gender.

Discussion

Utilizing the inner capabilities of an open source software is advantages in a low resource setting. We did such an exercise with ODK software by incorporating study subject selection algorithm. As epidemiological surveys are time-consuming on its data collecting part, this expedites the subject selection and data capture process through an open source electronic mobile application is an eminent choice. We adopted MySQL database for storage at the server after it gets synchronized during a favorable network time. ODK Aggregate is the data management platform adopted for the purpose which has installed on the institutional server. ODK aggregate allows export of data either CSV, JSON & KML format for analysis. We utilized visualize feature of ODK aggregate for data monitoring periodically guided by the instructions given in ODK program documents. This besides technical advantage, using ODK assisted algorithm epidemiological survey would overcome the difficulties of using pen and paper.[11] Though technical expertise in designing and handing the ODK program won’t be required to and expertise level, there one instance of hurdles we did face during the development. The is especially with respect server installation, the configuration of map visualization function at ODK Aggregate. At the advert usage of smartphone access the professionals, training the ODK android application faced minor obstacles.
Limitations

As a form utility, while incorporating changes in the application raised from the feedback on change of questions, we experienced certain issues in generating the updated application. Even a minor change like spell check on form mandates complete XML generation and entire updating of the existing application. Further uploading of modified XML in the server is treated as a new application. This insisted us to remove the existing application both at the server end and at frontend device which is little time-consuming. Also, as per the Kish method maximum of 6 eligible adults can be considered in the selection process. The survey decided for 20 households in a cluster.

Conclusion

Kish selection grid algorithm for subject selection was implemented using open source ODK application. This would facilitate the field investigator to select the eligible adult in each household. The application would reduce the time taken for completing the interview of study participants. ODK is an enterprise architecture, which is identified to develop a mobile application for a community based public health survey. A number of innovative solutions for data collection were introduced using android based data collection, an alternative for paper-based data collection, which includes open source data management system and electronic subject selection algorithm. Further studies over the application of open source platforms in key areas of public health such as disease surveillance and program management would address its wider usage in a low resource setting.

Summary table

<table>
<thead>
<tr>
<th>What was already known about the mobile application for an epidemiological survey?</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Electronic tools reduce the time duration of data collection in epidemiological surveys</td>
</tr>
<tr>
<td>- Using mobile application in epidemiological survey improves the quality of data</td>
</tr>
<tr>
<td>- Open source application could be a cost-effective and alternative to paper-based data collection in a low resource setting</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What this study added new knowledge in using mobile application for an epidemiological survey?</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Open source application has the capability for collecting data in large-scale national level epidemiological survey</td>
</tr>
<tr>
<td>- Open source application can be exploited to administer complex algorithms used in data collection in epidemiological surveys</td>
</tr>
<tr>
<td>- Subject selection algorithm like Kish methodology could be successfully implemented in open source mobile application for epidemiological surveys</td>
</tr>
</tbody>
</table>
Conflict of Interest: None

Funding source: This activity was undertaken for the National Non-communicable disease Monitoring Survey 2017-2018 funded by Indian Council of Medical Research, Department of Health Research, Government of India.
References

1. Kennedy JM, Vargus B. Challenges in survey research and their implications for philanthropic

   10;8(4):e2654.

3. El Emam K, Jonker E, Sampson M, Krleža-Jerić K, Neisa A. The Use of Electronic Data Capture
   Tools in Clinical Trials: Web-Survey of 259 Canadian Trials. J Med Internet Res [Internet]. 2009

4. Fitzgerald G, FitzGibbon M. A Comparative Analysis of Traditional and Digital Data Collection
   Methods in Social Research in LDCs - Case Studies Exploring Implications for Participation,

5. Anokwa Y, Hartung C, Brunette W, Borriello G, Lerer A. Open Source Data Collection in the


8. Extensible Markup Language (XML) 1.0 (Fifth Edition) [Internet]. [cited 2018 Mar 9]. Available
   from: https://www.w3.org/TR/xml/

9. Kumar RV. Respondent Selection Methods in Household Surveys [Internet]. Rochester, NY:
   Available from: https://papers.ssm.com/abstract=2392928


    questionnaires with PDA for behavioral surveys in Africa: Findings from a behavioral monitoring