Title: Electronic data capture for large scale typhoid surveillance, household contact tracing, and health utilisation survey: Strategic Typhoid Alliance across Africa and Asia.

Authors: Deus Thindwa[1,2], Yama G Farooq[3], Mila Shakya[4], Nirod Saha[5], Susan Tonks[3], Yaw Anokwa[6], Melita A Gordon[1,7], Carl Hartung[6], James E Meiring[3], Andrew J Pollard[3], Robert S Heyderman[1,8], on behalf of The Strategic Typhoid alliance across Africa and Asia consortium.

Affiliations

[1] Malawi Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi.

[2] Centre for the Mathematical Modelling of Infectious Diseases, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK.

[3] Centre for Clinical Vaccinology and Tropical Medicine, Department of Paediatrics, University of Oxford and the National Institute for Health, Oxford, UK.

[4] Oxford University Clinical Research Unit-Patan Academy of Health Sciences, Nepal.

[5] International Centre for Diarrhoeal Diseases Research, Dhaka, Bangladesh.

[6] Nafundi, Seattle, Washington, United States of America.

[7] Institute of Infection and Global Health, University of Liverpool, Liverpool, UK.

[8] Division of Infection and Immunity, University College London, London, UK.

* Corresponding author

Email: deus.thindwa@gmail.com (DT)

^ Membership of the Strategic Typhoid alliance across Africa and Asia consortium is provided in the Acknowledgements.

Abstract

Background: Electronic data capture systems (EDCs) have the potential to achieve efficiency and quality in collection of multisite data. We quantify volume, time, accuracy and costs of an EDC using large-scale census data from the STRATAA consortium, a comprehensive programme assessing population dynamics and epidemiology of typhoid fever in Malawi, Nepal and Bangladesh to inform vaccine and public health interventions.

Results: A census form was developed through a structured iterative process and implemented using Open Data Kit Collect running on Android-based tablets. Data were uploaded to Open Data Kit Aggregate, then auto-synced to MySQL-defined database nightly. Data were backed-up daily from 3 sites centrally, and auto-reported weekly. Pre-census materials' costs were estimated. Demographics of 308,348 individuals from 80,851 households were recorded within average of 14.7 weeks range (13-16) using 65 fieldworkers. Overall, 21.7 errors (95% confidence interval: 21.4, 22.0) per 10,000 data points were found: 13.0 (95% confidence interval: 12.6, 13.5) and 24.5 (95% confidence interval: 24.1, 24.9) errors on numeric and text fields respectively. These values meet standard quality threshold of 50 errors per 10,000 data points. The EDC's total variable cost was estimated at US$13,791.82 per site.

Conclusions: In conclusion, the EDC is robust, allowing for timely and high volume accurate data collection, and could be adopted in similar epidemiological settings.

Keywords: Africa, Asia, Electronic data capture, Open Data Kit, Typhoid fever.

Background

Use of electronic data capture systems (EDCs) for health research has increased since Apple's launch of the first handheld device in 1993 [1], and for observational studies and clinical trials is beginning to replace paper-based data collection methods. Paper-based systems have the advantage that they provide a hard copy source document but are characterised by high inaccuracies, substantial omissions, longer data turnaround time, longer data entry time, and high incremental costs both during the data collection and subsequent entry into an electronic database [2–6]. The advantages of EDC include built-in global positioning system (GPS) locator that automatically capture geographical coordinates thus minimizing transcription errors from external GPS locators; password-locked tablets and data encryption that maintain participant data confidentiality; required checks that prevent data omissions; range checks and data type checks that prevent typographical errors; skip patterns that provide logical responses; barcode technology that automates entry of unique identification; timestamps that provide a means to monitor work rate; and internet connectivity that ensures availability of real-time data [3, 4, 7, 8]. Despite these benefits [9], there is limited description of the performance of EDCs for large-scale or multisite surveys in low and middle-income countries.

Each year, an estimated 9.9-24.2 million typhoid fever cases occur from low and middle income countries resulting in approximately 75,000-208,000 deaths [10, 11]. However, although essential to build a public health case for disease control efforts such as vaccination and provision of clean water, sanitation and hygiene, obtaining reliable estimates for the burden of disease at national and sub-national level is difficult [12]. This requires collection of high quality field demographic, mapping, epidemiological, and clinical and laboratory data at scale from both hospital and community-based survey studies [13]. Interestingly, the

collection of such quality data is hindered by complexities of dilapidated health facilities, overcrowding, unstructured housing or slums, and illiteracy [14].

We present an open source-based EDC, designed to overcome data quality complexities, and evaluate the efficiency, quality, and costs of the EDC by measuring volume, time, accuracy, and material costs using multisite census data collected from sub-Saharan Africa and Asia. The EDC was developed and implemented within the Strategic Typhoid alliance across Africa and Asia (STRATAA), a comprehensive programme which is assessing population dynamics and epidemiology of typhoid fever in Malawi, Bangladesh and Nepal to inform design of vaccine and public health interventions.

Design, development and implementation

Study setting and participants

The census component of the STRATAA study aimed to collect demographics from approximately 100,000 individuals, of all ages, in each of the three sites, to form the sampling frame for subsequent sub-studies. More details of the STRATAA study design and participants have previously been described [13]. In brief, the three sites, one in each country, were selected based on high known burden of enteric fever, differing epidemiological patterns and previous ability to deliver paper-based studies of high participant volume and logistical complexity.

Electronic census report form and data cleaning procedure

An electronic census report form (eCRF), uniform to all sites, was developed through a structured iterative process. An eCRF comprised household- and individual-level questions. The eCRF data fields reflected a range of data types including integers to capture census team identifier, interviewer identifier, phone numbers of key respondent and older household members, household member number, and age; decimal to capture GPS points; alphanumeric to capture household unique identifier (barcode); texts to capture ward/traditional authority name, community/district name, physical address, respondent name, respondent relationship to head of household, respondent position in the household, head of household name, household member name, household member tribe/ethnicity, household member relationship to head, marital status, spouse name, education levels, employment status, mother's name, and father's name; characters to capture study site, household occupancy status, consent status, study information access status, sex, and school attendance status; and dates to capture household visit date and date of birth of each household member.

To ensure ultimate generation of error-free data , the eCRF data fields were designed with quality control tools such as dropdown menus, range checks, choice fields, skip patterns, required checks, double-data entry checks, systematic auto-numbering, preloading, and looping. However, due to other internal and external limitations of the EDC, we further built external database queries based on the Structured Query Language (SQL) to track potential data entry errors that might have arisen beyond EDC's control. External SQL queries were aimed to expose persistent error sources which included duplication of study household identifiers (barcode); duplication of entire individual demographics; barcode decoding errors during scan; illogical ages or date of births of children relative to parents; incorrect household visit dates relative to tablet system date; misspellings of traditional authority names/ward numbers, physical addresses, respondent names, and household members names; missing GPS points; inaccurate GPS points relative to the household; and mismatches between community names and GPS points. After running the external SQL queries on the census database table and identifying the errors, each correction of an error by the data officer triggered an automatic log to an audit-trail table with entries (table's column names) that included table name with error, action on an error (update, insertion, or deletion), individual/household barcode identifier with an error, field name with an error, old value, new value, timestamp, and a user's name modifying an error. This generated a single row in an audit-trail table for each single error that was modified in the original census table. Errors corresponding to GPS points were specifically identified through sub-setting and importing GPS points (longitude, latitude, and altitude) from the census table into Google Earth Pro software (Google LLC, Mountain View, California, USA) as a keyhole markup language file, and then mapping the GPS points on the overlay of community boundaries' and households' satellite images. Once a GPS point was not mapped within 5 meters at 10% accuracy of the household or within the community boundary, it was considered a mapping error, and

corrected through remapping in the field and updating it in the census table thereby triggering an audit-trail table error record. All the other errors exposed by the external SQL queries were investigated thoroughly in the field before corrections could be applied to the census table and subsequently auto-logged into the audit-trail table. The maximum number of visits to the household prior declaring the household vacant or errors permanently unresolved was twice. We provide the flow diagram of the eCRF in (Fig. 1), whereas the technical details of the extensible markup language code used to create an eCRF, and the SQL code used to create the audit-trail table and triggers to the audit-trail table have been publicly shared through GitHub (GitHub Inc, San Francisco, California, USA) [15].

Figure 1. Electronic census report form (eCRF) flowchart.

Electronic data capture system

We designed a uniform EDC using combined open-source tools; Open Data Kit (ODK) software (Nafundi, Seattle, Washington, USA) [16–18], and MySQL relational database management system (Oracle Corporation, Redwood city, California, USA) [19]. The eCRF was customized in ODK Collect and uploaded onto Android-based Asus ZenPad (AsusTek Computer Inc., Taipei, Taiwan), and Samsung (Samsung group, Seoul, South Korea) tablets. Then data were collected in the field during the day and temporarily saved in the tablet's memory. At the end of each day, tablets were returned to the base STRATAA data office and data were uploaded from the tablet's memory to ODK Aggregate server via a secure wireless network technology. Tablets were then charged overnight at the base data office in preparation for use on the next day. For every scheduled time of the night, data automatically synchronized from ODK Aggregate server to MySQL-defined database, set up for four main reasons; first, to facilitate corrections of inconsistencies beyond ODK validations (e.g. all

persistent error sources mentioned above) and auto-audit the corrections; second, to ensure homogeneous database structure across sites in order to facilitate multisite dataset merging, and to preserve meaningful variables (excluding metadata generated by ODK software) in order to provide intuitive datasets to epidemiologists and statisticians; third, to generate automated reports using SQL; and last, to allow automated back-up of cleansed data from MySQL-defined database to external storage devices. The EDC also allows daily comma-separated value and anonymized data format to securely and automatically synchronize from each site's ODK Aggregate server to a central repository. Conversely, the comma-separated value data format, from MySQL-defined database, are sporadically exported back to tablet's ODK media folder to enable data preloading for sub-sequent sub-studies (Fig. 2). Technical details of the scripts for synchronizations, and creation of table structures and triggers have been publicly shared through GitHub [15].

Figure 2. Electronic data capture system for a multisite study. MySQL-defined databases b_strataa, k_strataa, and d_strataa have homogeneous structures (*) e.g. table columns, data types, triggers or views. Data from MySQL-defined database table are exported back to Android-based tablet enabling data preloading for subsequent sub-studies (P). Homogeneous databases across sites merge enabling multisite data analyses (H).

Pre-census time, costs, and training

We estimated time and costs required to attain the following census-related materials or complete census activities; tablets (including screen protectors, and protective covers), desktop server computers, network devices, barcodes, development of eCRF, training of field workers, replacement of broken tablets, and backpacks. We did not assess other operational costs because of uncertainty e.g. electric power to servers, charging tablets, and electronic

data synchronization. We trained fieldworkers and assessed their suitability to conduct census by administering a practical mock test and then selecting best performers. Moreover, five weeks post-census implementation, we retrained fieldworkers based on calculated individual performances on data quality and data collection speed.

Statistical analysis and visualization

We estimated the error rates, after running external SQL queries but prior to data cleaning, by dividing the total number of errors observed by the total number of data points (≈ all expected errors). A data point was defined as a discrete unit of information that could possibly be obtained from each member of the population after administering an eCRF e.g. If an eCRF had (n) number of unique questions, with each question corresponding to a variable (X_i), for (N) number of respondents, then the total data points for eCRF would be $\sum_{i=1}^{n}(X_i N)$. In our calculations, data points for household and individual-level variables were calculated separately and summed up. The reason was that household-level questions were answered by a key informant (head of household or respondent ≥ 18 years old) while individual-level questions were hypothetically answered by all household members (represented by a key informant). Exact binomial confidence intervals were used to estimate error rates. Data entry speed and accuracy by fieldworkers were combined into a single merit in order to measure their performance [20]. For each fieldworker, we standardized the data entry speeds (z_s) and errors (z_e), and assigned more weight to data entry speed (60%) than errors (40%) given the background that the EDC was robustly developed to prevent most data entry errors, thus, speed was more important. The final data entry speed-accuracy trade-off was calculated using the formula $(SAT = -z_s * 0.6 - z_e * 0.4)$ where $z_s = (s - \mu_s)/\delta_s$ and $z_e = (e - \mu_e)/\delta_e$, (s) is the total speed for all data entries per field worker, (μ_s) is the mean speed for all fieldworkers, (δ_s) is the speed standard deviation, (e) is the total number of errors per field

worker, (μ_e) is the mean error for all fieldworkers, and (δ_e) is the error standard deviation. In addition, we used Wilcoxon Signed-Rank Test for paired samples pre- versus post-retraining in order to measure any statistical difference in the number of errors committed, and determine whether retraining the fieldworkers helped improve accuracy. All statistics and plots were conducted in R version 3.4.0 [21], eCRF flowchart and EDC diagram were created using www.draw.io (JGraph, London, England).

Results

Data collection volume, time and accuracy

We recorded demographics of 308,348 individuals from 80,851 households in three countries between June 2016 and October 2016; 97,410 individuals and 22,364 households from Malawi, 100,207 and 32,368 from Nepal, and 110,731 and 26,119 from Bangladesh. Completeness of household demographics enumeration were 94.2%, 75.6% and 79.2% for Malawi, Nepal and Bangladesh, respectively, relative to background household count. The average number of weeks for enumeration was 14.7 range (13-16) using 20, 25, 20 field workers from Malawi, Nepal and Bangladesh, respectively. Overall, 21.7 errors (95% confidence interval: 21.4, 22.0) per 10,000 data points were found; 15.9 errors (95% confidence interval: 15.4, 16.4), 34.2 errors (95% confidence interval: 33.5, 34.9), and 14.6 errors (95% confidence interval: 14.2, 15.0) per 10,000 data points from Malawi, Nepal and Bangladesh, respectively. Of the 17,707 errors documented from all sites, the majority 12,740 (72.0%) occurred on text fields compared to numeric fields 3,868 (21.8%). In addition, 1,099 (6.2%) errors occurred as duplicate records (e.g. either by enumerating a household or any of its members at least twice) (Table 1).

Of all the data entry errors observed during enumeration period, 2,611 (65.4%), 6,265 (65.8%) and 3,013 (71.8%) were, respectively, committed in Malawi, Nepal and Bangladesh prior fieldworkers' retraining. Moreover, there were fewer errors observed after retraining of fieldworkers compared to pre-retraining, and the differences were statistically significant in Malawi (W =5.5, P <0.001), Nepal (W =19.5, P <0.001), and Bangladesh (W =0, P <0.001) (Fig. 3).

Figure 3. Data entry errors before and after retraining of fieldworkers, 2016.

Table 1. Census Data Collection Time, Volume and Accuracy in Three Typhoid Endemic Sites, 2016.

Study site	Time period of data collection	Total households	Total individuals	Number of errors*	Number of data points	Errors per 10,000 data points	95% CI**
All sites							
Overall	14.7 weeks (13-16)	80,851	308,348	17,707	8,173,179	21.7	21.4, 22.0
Numeric	14.7 weeks (13-16)	80,851	308,348	3,868	2,966,946	13.0	12.6, 13.5
Text	14.7 weeks (13-16)	80,851	308,348	12,740	5,206,233	24.5	24.1, 24.9
Malawi[§]							
Overall	Jul 2016 – Oct 2016	22,364	97, 410	3,991	2,515,254	15.9	15.4, 16.4
Numeric[Ŧ]	Jul 2016 – Oct 2016	22,364	97, 410	900	905,510	9.9	9.3, 10.6
Text[□]	Jul 2016 – Oct 2016	22,364	97, 410	2,291	1,609,744	14.2	13.7, 14.8
Nepal[§]							
Overall	May 2016 – Sep 2016	32, 368	100, 207	9,522	2,784,075	34.2	33.5, 34.9
Numeric[Ŧ]	May 2016 – Sep 2016	32, 368	100, 207	2,171	1,025,129	21.2	20.3, 22.1
Text[□]	May 2016 – Sep 2016	32, 368	100, 207	7,131	1,758,946	40.5	39.6, 41.5

Bangladesh§							
Overall	Jun 2016 – Aug 2016	26,119	110,731	4,194	2,873,850	14.6	14.2, 15.0
Numeric₮	Jun 2016 – Aug 2016	26,119	110,731	797	1,036,307	7.7	7.2, 8.23
Text□	Jun 2016 – Aug 2016	26,119	110,731	3,318	1,837,543	18.1	17.5, 18.7

* Persistent error sources included duplication of household identifiers (barcodes); duplication of entire individual demographics; incorrect barcode decoding during scan; illogical ages or date of births of children relative to parents; incorrect household visit dates relative to tablet system date; misspellings of traditional authority names/ward numbers, physical addresses, respondent names, household members' names; missing GPS points; inaccurate GPS points relative to the household; and mismatches between community names and GPS points. Duplicates resulted in 800 records being deleted in Malawi, 220 in Nepal, and 79 in Bangladesh.

₮ Includes numeric integer, numeric decimal and alphanumeric (barcode) data types.

□ Includes text, character, and date data types.

§ Number of census field workers for Malawi (20), Nepal (25), and Bangladesh (20).

** CI: Confidence Interval estimated by binomial (Clopper-Pearson) 'exact' method based on the error distribution.

Time and cost of census materials

The time required to attain each material or complete each activity in preparation for census implementation varied by study site, ranging from 2 to 60 days. The most time-consuming activity was the development and customization of eCRF, which was completed in 60 days collectively. This was followed by the procurement of tablets and backpacks, which were acquired in between 7 and 60 days. In addition, we also procured and designed household identifier (barcode) stickers in between 7 and 21 days. Replacement of malfunctioned tablets reported by each study was accomplished within 30 days. We extensively trained our study fieldworkers for up to 5 days focussing on the study protocol, practical aspect of completing an eCRF, and community engagement skills. Selection of potential fieldworkers to join the study team was sorely based on successful completion of the training. Computer servers and network devices to enable data storage and transfers from tablets were pre-existing in Malawi and Bangladesh, and newly acquired in Nepal within 30 days (Table 2).

The major variable cost was incurred by customization of eCRF for use in ODK Collect for a total of US$ 9,000 for all sites, followed by procurement of 27 tablets at a variable cost of US$5,407.02. Other prominent variable costs included procurement of a desktop server (at US$1,523.21), training 27 field workers to use an eCRF and in field practices (at US$1,479.60), procurement and shipment of 27 backpacks (at US$1,277.91) and 1,500 barcode sheets (at US$720.00), replacement of a malfunctioned tablet (at US$200.26) and procurement of a network router (at $183.82). The total variable cost for the EDC was US$13,791.82 per site (Table 2).

Table 2. Time and Costs Attainment Prior to Implementation of an Electronic Data Capture System (EDC) in Three Typhoid Endemic Sites, 2016.

Material or activity**	Time to attain item or complete activity varied by site		Number of units required (Range)	Unit cost (US$)*	Variable cost (US$)
Category	Days	Unit	$X_1 – X_2$	Y	$X_1 . Y$
Tablets (including screen protectors and protective cover) §	7 – 60	Tablet	27 – 42	200.26	5,407.02
Desktop server computers §	0 – 30	Computer	1 – 4	1,523.21	1,523.21
Network devices §	0 – 30	Router	1 – 4	183.82	183.82
Barcodes	7 – 21	Sheet	1,500 – 2,530	0.48	720.00
Electronic census report form (eCRF) development and customization Ŧ	60	eCRF	1 – 3	3,000.00	3,000.00
Training field workers	2 – 5	Field worker	27 – 37	56.82	1,479.60
Replacement of malfunctioned tablets	7 – 30	Tablet	1 – 3	200.26	200.26
Backpacks	7 – 60	Backpack	27 – 42	47.33	1,277.91

* Average unit cost estimated in 2016 across all study sites.

Ŧ Only 1 uniform eCRF was developed for 3 sites, for purposes of calculations, we divide the total cost by 3.

§ Some tablets already existed in other sites. Similarly, network devices and computer servers pre-existed in Malawi, Bangladesh, and a central coordinating site (Oxford Vaccine Group) but not in Nepal.

287 ** Excludes costs of electric power to servers, charging tablets and data synchronization because of uncertainty.

288 US$ United States dollar currency.

289

Discussion

In this study, we have developed and implemented an EDC which allows high volume of data collection over short time periods, high data accuracy, 12-hourly updated data access, and quality checking for decision making. Additionally, the EDC is robust, allowing for automated reports generation, scalability and could be adaptable to other epidemiological settings. Finally, the total variable cost of the EDC's pre-census materials and activities, was minimal relative to paper-based data collection methods from similar settings.

Data were collected by largely secondary school level only fieldworkers receiving 1 week of training and a day of retraining, and although the learning curve of using an eCRF in ODK Collect on Android-based tablets was steep in the first 5 weeks of field work, high volume and fairly accurate data were recorded (Fig. 3 and Fig. 4). The data accuracy of ~0.22% errors (21.7 errors per 10,000 data points) reported in this study meets the acceptable quality threshold of 50 errors per 10,000 data points recommended by the Society of Clinical Data Management (SCDM, McLean, Virginia, USA) [22, 23]. The highly accurate EDC data in this study is comparable to EDC data accuracies reported by the chronic disease research in South Africa (0.17%) and maternal health survey in Burkina Faso (0.24%) [24, 25]. However, our EDC data accuracy is superior to EDC data accuracies reported by the maternal health (2.8%) and neglected tropic disease surveys (5.2%) in Ethiopia, the bloodstream infections study in Zanzibar (1.0%) and the tuberculosis program in India (4.2%) [2, 3, 7, 26]. Moreover, our EDC data are more accurate in comparison to data reported from paper-based studies of maternal health (1.1%) and neglected tropical disease (6.2%) surveys in Ethiopia, bloodstream infections study in Zanzibar (7.0%), chronic disease research in South Africa (0.73%), and randomized controlled trial in Fiji (20.8%) [2–4, 7, 24]. As with previous studies [2, 22, 27], text fields of this eCRF generated more errors than numeric fields, , and

suggest that such errors could be prevented in eCRF designs by minimizing the use of text fields through coding of text responses or leaving out insignificant text responses completely. The accuracy variations between EDCs are probably due to robustness of the EDC design in terms of error proofing. Robustness in the design is likely to depend on the limitations of software and hardware, and technical know-how of developers.

Figure 4. Speed and accuracy trade-off before and after retraining of fieldworkers, 2016.

Unlike the EDC and paper-based methods used in a similarly setting [28], our EDC synchronized study data updates at least every 12 hours post-data collection in order to provide recent data accessibility for decision making; Rapid accessibility to recent data has enabled immediate quality checks and data cleaning on critical variables which, at the time of the study, are beyond ODK's built-in validations. It also enabled us to quickly understand and decide on ways to improve participant uptake rates, adding to a growing body of literature reporting how rapid data updates by an EDC enables swift decisions [9, 29, 30].

The EDC was also designed to counteract some complexities associated with data collection in low- and middle-income; Internet connectivity was through a client-server system where data capture client (ODK Collect) was an offline stand-alone instance separated from the database server (ODK Aggregate). Data were synchronized from client to server at a later point in time at the base STRATAA data office where connectivity was possible. This approach has also been recommended by others [31, 32], and we did not experience any damage or theft of the tablets which led to data loss before data was synchronized to database server. We adhered to a practice of disabling eCRF 'edit' options, post-interview, in order to maintain data integrity in the field. Validations within the ODK Collect prevented most

errors. However, 0.4% duplicate household identities and 0.3% missing GPS points were uncovered in addition to other text and numeric errors. Following good data management practices [33], our EDC also provided three backup strategies; scheduled data synchronization to (i) centralized repository, (ii) MySQL-defined databases, and (iii) scheduled incremental backup of MySQL-defined databases to external storage devices.

The EDC delivered considerable capacity for automated report generation, scalability and adaptability. We were able to use SQL to pull seasonal data from MySQL-defined database, and automate summaries of demographics in order to monitor progress of field work, and collective and individual performance of field workers. SQL was preferred because of its simple but powerful syntax, and its wider use in handling complex queries to epidemiological datasets [22, 24, 29, 32]. Since the STRATAA consortium continuously generates laboratory data, post-census, the EDC also allows scalability, pushing laboratory data from laboratory database systems to MySQL-defined databases while keeping the database structure homogeneous across sites. The EDC could therefore not only be adopted by others collecting large data volumes requiring centralized data storage and automation of process, but also be tested by settings with little experience in conducting field-based research. The EDC is installed in three typhoid endemic settings and will be maintained by STRATAA consortium for adaptability of potential future studies.

Costs estimates on the data capture systems across low- and middle-income settings account for different item inclusions [7, 28, 34, 35]. However, generally, our total variable cost of the EDC was minimal relative to most EDCs or paper-based data collection methods conducted in similar settings. For instance, our EDC's total variable cost is analogous to US$13,883.00 incurred on a paper-based survey of neglected tropic diseases in Ethiopia [7]. However, in

365 northern Malawi, estimated total variable costs of an EDC (US$14,477.46 [£11,427]) and
366 paper-based system (US$23,939.06 [£18,895]) are slightly and much higher than our EDC
367 respectively [28]. Similarly, our total variable cost is relatively low compared to paper-based
368 studies conducted in Bangladesh and Philippines (US$45,000.00) on verbal autopsy [34], and
369 in Kenya (US$15,999.00) on influenza [35].
370

371 # Conclusion

372 In conclusion, we have designed an EDC which has been implemented in three typhoid
373 endemic sites to collect large volume of accurate data in short time periods with rapid access
374 through automated reports. The EDC's development required careful attention to detail but
375 the materials' variable costs prior to census implementation, were minimal relative to some
376 EDCs and paper-based data collection methods. This EDC could be adopted in similar
377 epidemiological settings, enabling the collection and management of large data volumes,
378 centralize data storage, and automated data processes.
379

Availability and requirements

The code scripts used to develop the EDC (ODK Collect eCRF and MySQL database objects), and the raw data for errors analysed in this paper are all available through GitHub [15].

Abbreviations

EDCs: Electronic data capture systems; STRATAA: Strategic Typhoid alliance across Africa and Asia consortium; ODK: Open Data Kit; GPS: global positioning system ; eCRF: electronic census report form; SQL: Structured Query Language; CI: Confidence Intervals; US$: United States dollar; SCDM: Society of Clinical Data Management.

Declarations

Ethics approval and consent to participate

Ethical approval was obtained from the Malawi National Health Sciences Research Committee, 15/5/1599; Bangladesh ICDDR,B Institutional Review Board, PR-15119; Nepal Health Research Council, 306/2015; and Oxford Tropical Research Ethics Committee, 39-15. Following extensive sensitisation and engagement with community and traditional leaders, and community health-workers, the key informant from each household provided a verbal informed consent, to enumerate the household, which was documented in the eCRF.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are available in GitHub [15].

Competing interests

The authors declare that they have no competing interests.

Funding

Funding for the STRATAA study has been provided by a Wellcome Trust Strategic Award (no. 106158/Z/14/Z), https://wellcome.ac.uk/funding/managing-grant/grantsawarded, and the Bill and Melinda Gates Foundation (no. 617 OPP1141321), https://www.gatesfoundation.org/How-We-Work/Quick-Links/Grants-Database to AJP. The Malawi-Liverpool-Wellcome Programme and the Oxford University Clinical Research Unit in Vietnam are supported by the Wellcome Trust with Major Overseas Programme core awards. The funders did not play any role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript.

Authors contribution

Conceptualization: DT, YGF, RSH; Methodology: DT, YGF, MS, NS, YA, CH; Writing-Original Draft: DT; Writing-Review and Editing: YGF, MS, NS, ST, YA, MAG, CH, JEM, AJP, RSH; Funding-Acquisition: MAG, AJP, RSH. All authors read and approved the final manuscript.

Acknowledgements

Members of The Strategic Typhoid alliance across Africa & Asia consortium (STRATAA): Oxford University Clinical Research Unit, Patan Academy of Health Sciences, Kathmandu, Nepal (Mila Shakya, Abhilasha Karkey, Sabina Dongol, Amit Aryjal, Buddha Basnyat);

International Center for Diarrhoeal Diseases Research, Dhaka, Bangladesh (Nirod Saha, Farhana Khanam, Md Arifuzzaman Khan, John D. Clemens, Firdausi Qadri, K. Zaman); Malawi Liverpool Wellcome Trust Clinical Research Programme, Malawi (Deus Thindwa, Robert S Heyderman, Melita A Gordon, Tikhala Makhaza Jere, Chisomo Msefula, Tonney Nyirenda); The Hospital for Tropical Diseases, Wellcome Trust Major Overseas Programme, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam (Tan Trinh Van); Wellcome Trust Sanger Institute, Cambridge, United Kingdom (Stephen Reece, Gordon Dougan); Oxford Vaccine Group, Department of Paediatrics, University of Oxford, and the NIHR Oxford Biomedical Research Centre, Oxford, United Kingdom (Merryn Voysey, Christoph J. Blohmke, Jennifer Hill, Thomas C. Darton, Susan Tonks, Yama G Farooq, James E. Meiring, Andrew J Pollard); Yale School of Public Health, Yale University, New Haven, Connecticut, United States of America (Neil J. Saad, Virginia E. Pitzer); Center for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, United Kingdom (Stephen Baker, Christiane Dolecek); The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Australia (Sarah J. Dunstan); Centre for Systems Genomics, University of Melbourne, Parkville, Victoria, Australia (Kathryn E. Holt). Division of Infection and Immunity, University College London (Robert S Heyderman), Department of Infectious Disease Epidemiology, Imperial College London (Deus Thindwa), Institute of Infection and Global Health, University of Liverpool, Liverpool, United Kingdom (Melita A Gordon). We are grateful to all individuals who participated in the census of the STRATAA study. We also thank all community leaders from the STRATAA areas for allowing us to conduct STRATAA study in their areas.

References

1. Garritty C, Emam KE. Who's Using PDAs? Estimates of PDA Use by Health Care Providers: A Systematic Review of Surveys. J Med Internet Res. 2006;8:e7.

2. Medhanyie AA, Spigt M, Yebyo H, Little A, Tadesse K, Dinant G-J, et al. Quality of routine health data collected by health workers using smartphone at primary health care in Ethiopia. Int J Med Inf. 2017;101:9–14.

3. Thriemer K, Ley B, Ame SM, Puri MK, Hashim R, Chang NY, et al. Replacing paper data collection forms with electronic data entry in the field: findings from a study of community-acquired bloodstream infections in Pemba, Zanzibar. BMC Res Notes. 2012;5:113.

4. Yu P, de Courten M, Pan E, Galea G, Pryor J. The development and evaluation of a PDA-based method for public health surveillance data collection in developing countries. Int J Med Inf. 2009;78:532–42.

5. Rorie DA, Flynn RWV, Grieve K, Doney A, Mackenzie I, MacDonald TM, et al. Electronic case report forms and electronic data capture within clinical trials and pharmacoepidemiology. Br J Clin Pharmacol. 83:1880–95.

6. Ali M, Deen JL, Khatib A, Enwere G, Seidlein L von, Reyburn R, et al. Paperless registration during survey enumerations and large oral cholera mass vaccination in zanzibar, the United republic of Tanzania. Bull World Health Organ. 2010;88:556–9.

7. King JD, Buolamwini J, Cromwell EA, Panfel A, Teferi T, Zerihun M, et al. A Novel Electronic Data Collection System for Large-Scale Surveys of Neglected Tropical Diseases. PLOS ONE. 2013;8:e74570.

8. King C, Hall J, Banda M, Beard J, Bird J, Kazembe P, et al. Electronic data capture in a rural African setting: evaluating experiences with different systems in Malawi. Glob Health Action. 2014;7:25878.

9. White A, Thomas DSK, Ezeanochie N, Bull S. Health Worker mHealth Utilization: A Systematic Review. Comput Inform Nurs CIN. 2016;34:206–13.

10. Antillón M, Warren JL, Crawford FW, Weinberger DM, Kürüm E, Pak GD, et al. The burden of typhoid fever in low- and middle-income countries: A meta-regression approach. PLoS Negl Trop Dis. 2017;11:e0005376.

11. Mogasale V, Maskery B, Ochiai RL, Lee JS, Mogasale VV, Ramani E, et al. Burden of typhoid fever in low-income and middle-income countries: a systematic, literature-based update with risk-factor adjustment. Lancet Glob Health. 2014;2:e570–80.

12. Crump JA. Building the case for wider use of typhoid vaccines. Vaccine. 2015;33 Suppl 3:C1–2.

13. Darton TC, Meiring JE, Tonks S, Khan MA, Khanam F, Shakya M, et al. The STRATAA study protocol: a programme to assess the burden of enteric fever in Bangladesh, Malawi and Nepal using prospective population census, passive surveillance, serological studies and healthcare utilisation surveys. BMJ Open. 2017;7:e016283.

14. MacPherson P, Choko AT, Webb EL, Thindwa D, Squire SB, Sambakunsi R, et al. Development and Validation of a Global Positioning System–based "Map Book" System for Categorizing Cluster Residency Status of Community Members Living in High-Density Urban Slums in Blantyre, Malawi. Am J Epidemiol. 2013;177:1143–7.

15. STRATAA consortium. Electronic Data Capture for Large Scale Typhoid Surveillance. 2017. https://github.com/Oxfordvaccinegroup/Electronic-Data-Capture-for-Large-Scale-Typhoid-Surveillance---STRATAA.

16. University of Washington. Open Data Kit. https://opendatakit.org/. Accessed 9 Jul 2017.

17. Anokwa Y, Hartung C, Brunette W, Borriello G, Lerer A. Open Source Data Collection in the Developing World. Computer. 2009;42:97–9.

18. Hartung C, Lerer A, Anokwa Y, Tseng C, Brunette W, Borriello G. 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. New York, NY, USA: ACM; 2010. p. 18:1–18:12. doi:10.1145/2369220.2369236.

19. Oracle Corporation, et al. MySQL. https://www.mysql.com/. Accessed 9 Jul 2017.

20. Chignell M, Tong T, Mizobuchi S, Delange T, Ho W, Walmsley W. Combining Multiple Measures into a Single Figure of Merit. Procedia Comput Sci. 2015;69:36–43.

21. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/. Accessed 28 May 2019.

22. Jenkins TM, Boyce TW, Akers R, Andringa J, Liu Y, Miller R, et al. Evaluation of a Teleform-based data collection system: A multi-center obesity research case study. Comput Biol Med. 2014;49:15–8.

23. Pomerantseva V, Ilicheva O. Clinical Data Collection, Cleaning and Verification in Anticipation of Database Lock. Pharm Med. 2011;25:223–33.

24. Dillon DG, Pirie F, Rice S, Pomilla C, Sandhu MS, Motala AA, et al. Open-source electronic data capture system offered increased accuracy and cost-effectiveness compared with paper methods in Africa. J Clin Epidemiol. 2014;67:1358–63.

25. Byass P, Hounton S, Ouédraogo M, Somé H, Diallo I, Fottrell E, et al. Direct data capture using hand-held computers in rural Burkina Faso: experiences, benefits and lessons learnt. Trop Med Int Health. 2008;13:25–30.

26. Patnaik S, Brunskill E, Thies W. Evaluating the Accuracy of Data Collection on Mobile Phones: A Study of Forms, SMS, and Voice. 2009. https://www.cs.cmu.edu/~ebrun/patnaik-ictd09.pdf. Accessed 22 May 2017.

27. Walther B, Hossin S, Townend J, Abernethy N, Parker D, Jeffries D. Comparison of Electronic Data Capture (EDC) with the Standard Data Capture Method for Clinical Trial Data. PLOS ONE. 2011;6:e25348.

28. McLean E, Dube A, Saul J, Branson K, Luhanga M, Mwiba O, et al. Implementing electronic data capture at a well-established health and demographic surveillance site in rural northern Malawi. Glob Health Action. 2017;10:1367162.

29. Rajput ZA, Mbugua S, Amadi D, Chepn□eno V, Saleem JJ, Anokwa Y, et al. Evaluation of an Android-based mHealth system for population surveillance in developing countries. J Am Med Inform Assoc JAMIA. 2012;19:655–9.

30. 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. doi:10.2196/mhealth.7827.

31. Meyer J, Fredrich D, Piegsa J, Habes M, van den Berg N, Hoffmann W. A mobile and asynchronous electronic data capture system for epidemiologic studies. Comput Methods Programs Biomed. 2013;110:369–79.

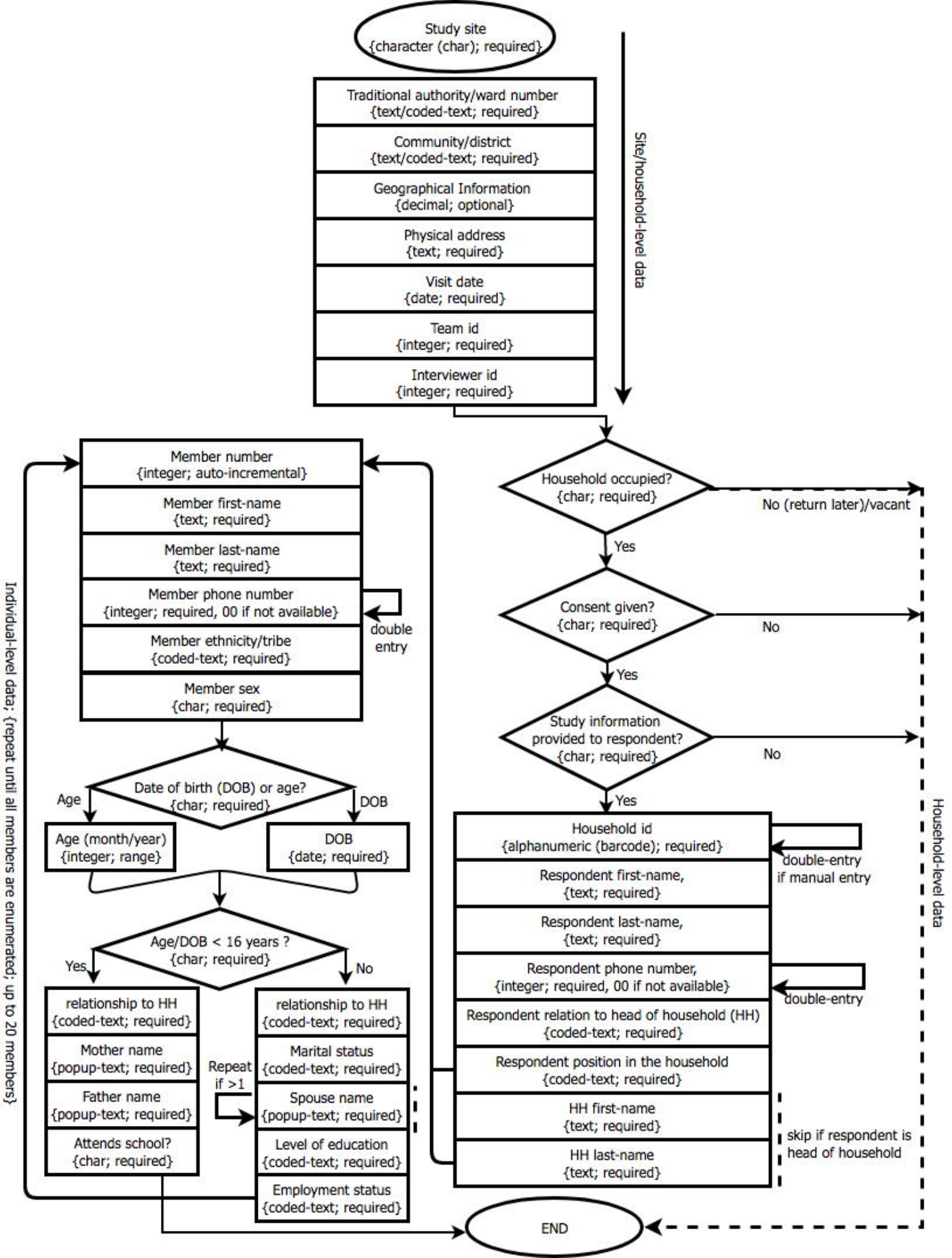
32. Baguiya A. An offline mobile data capture module for health and demographic surveillance system (HDSS) studies. Thesis. 2016. http://wiredspace.wits.ac.za/handle/10539/21399. Accessed 18 Jul 2017.

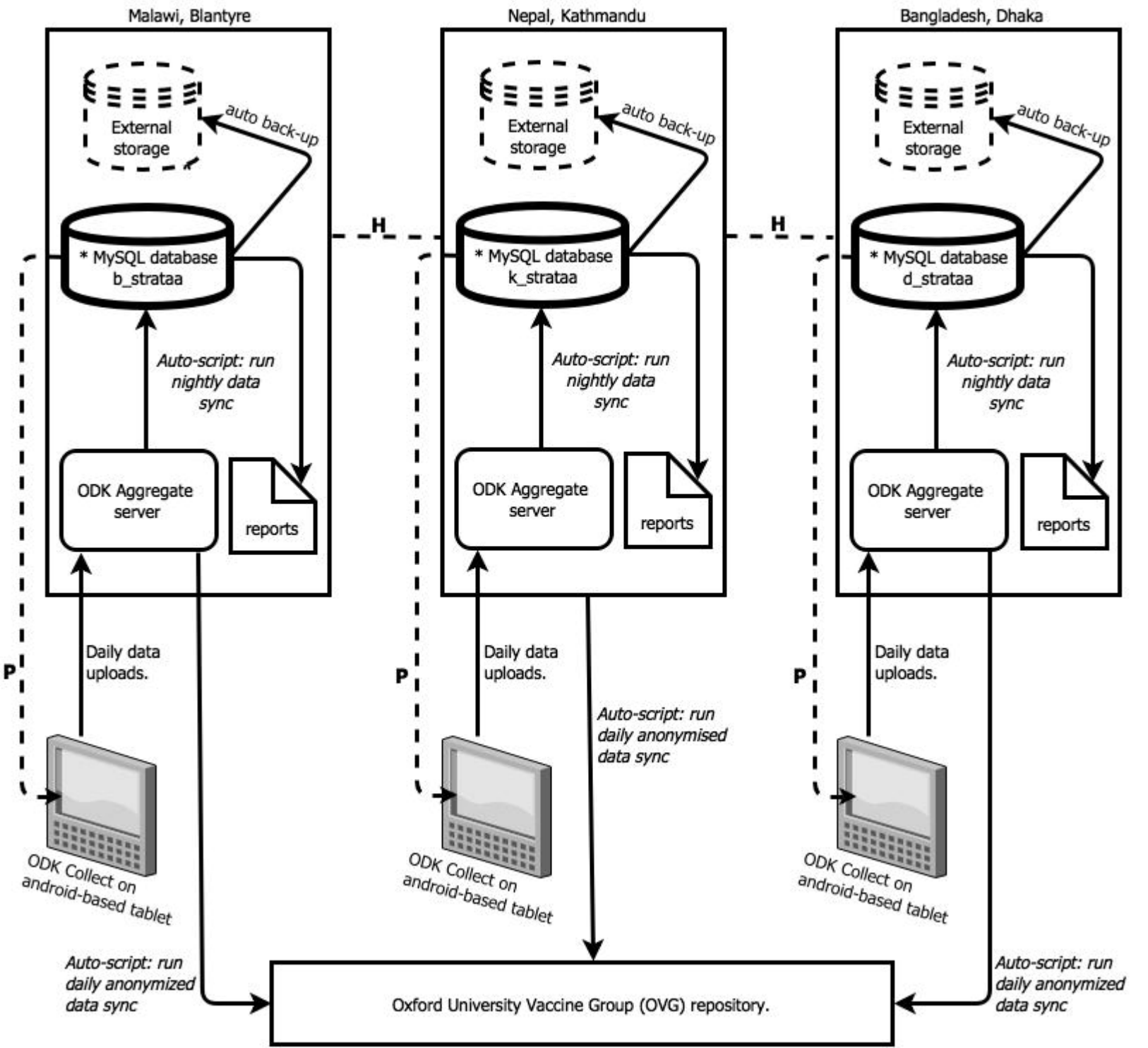
33. Shirima K, Mukasa O, Schellenberg JA, Manzi F, John D, Mushi A, et al. The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania. Emerg Themes Epidemiol. 2007;4:5.

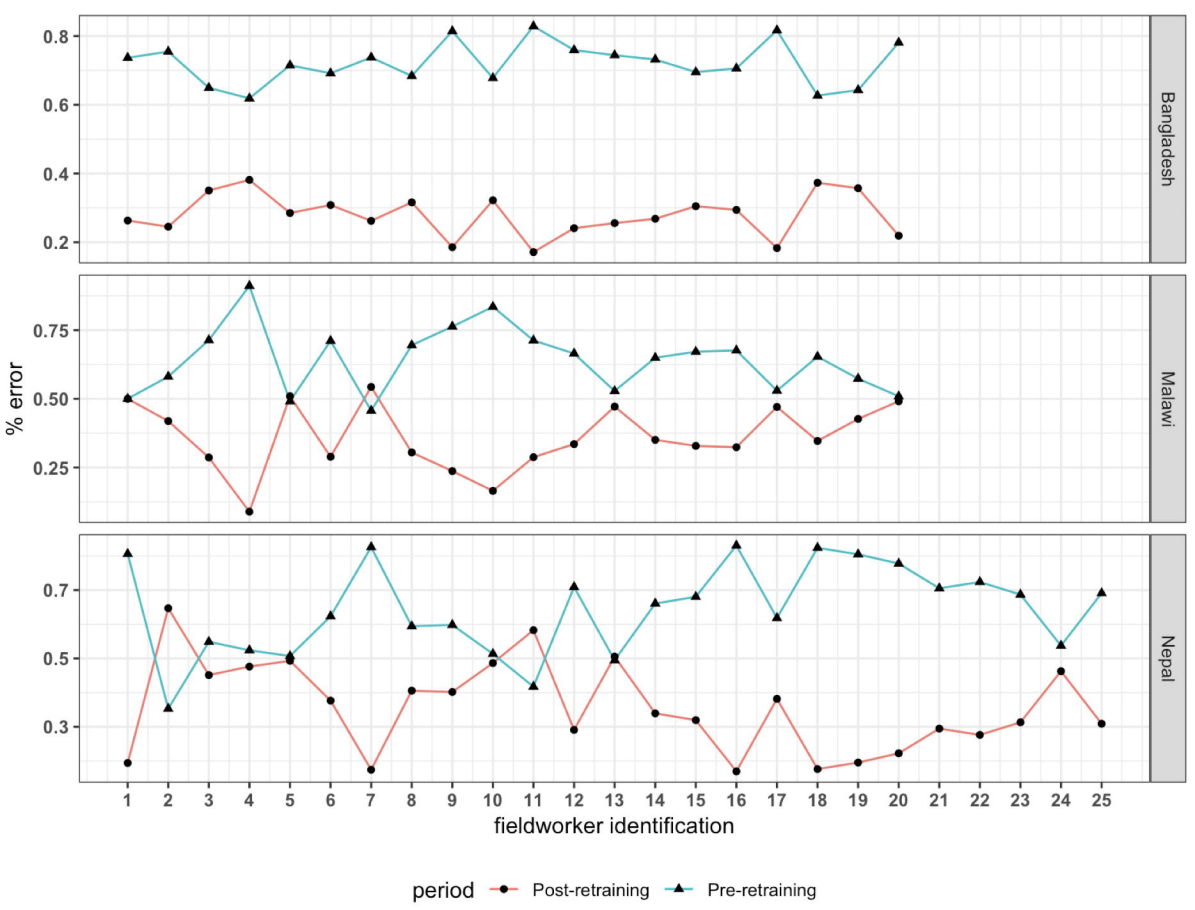
34. Flaxman AD, Stewart A, Joseph JC, Alam N, Alam SS, Chowdhury H, et al. Collecting verbal autopsies: improving and streamlining data collection processes using electronic tablets. Popul Health Metr. 2018;16:3.

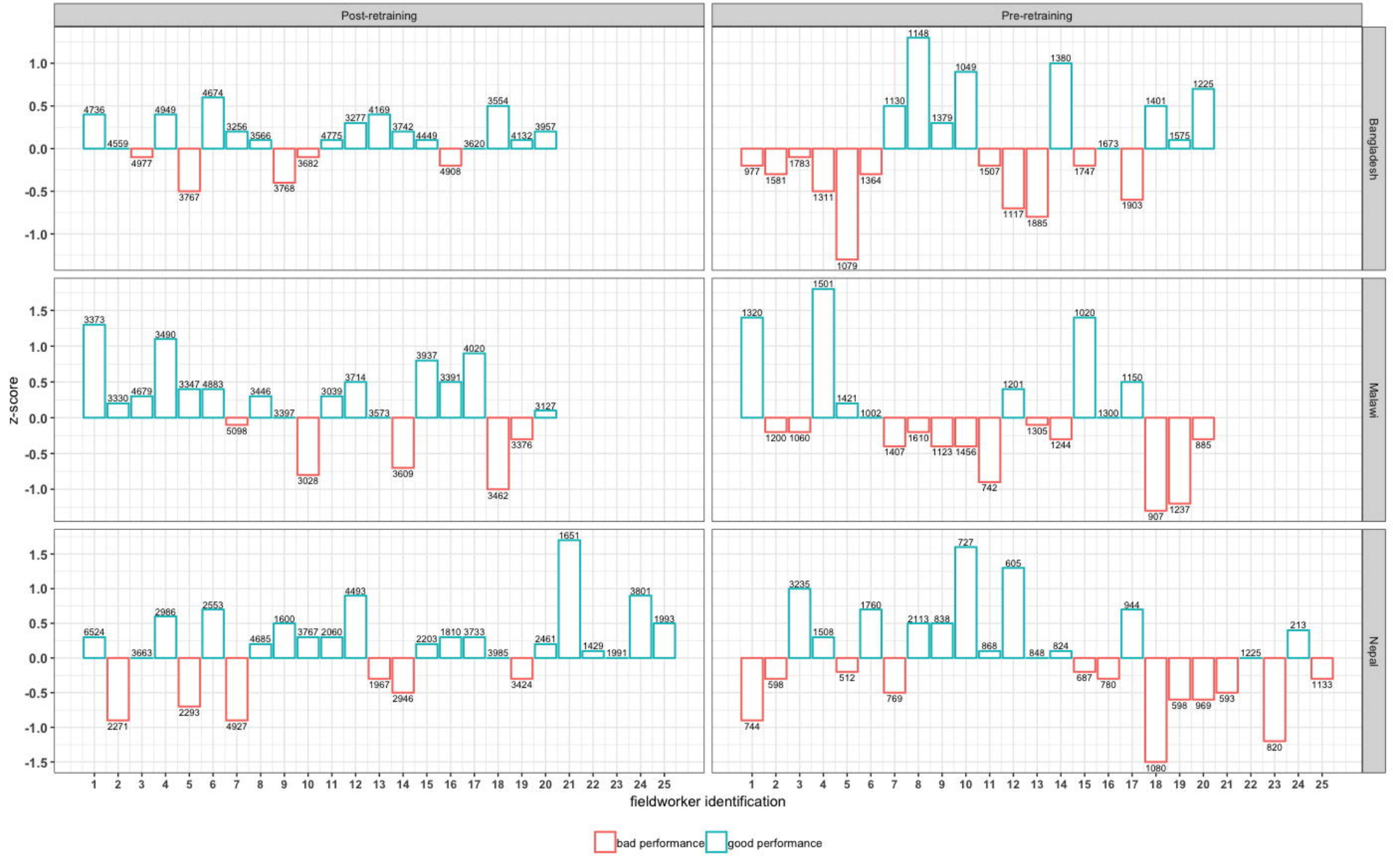
35. Njuguna HN, Caselton DL, Arunga GO, Emukule GO, Kinyanjui DK, Kalani RM, et al. A comparison of smartphones to paper-based questionnaires for routine influenza sentinel surveillance, Kenya, 2011–2012. BMC Med Inform Decis Mak. 2014;14:107.

550


Study site
{character (char); required}
Traditional authority/ward number
{text/coded-text; required}
Community/district
{text/coded-text; required}
Geographical Information
{decimal; optional}
Physical address
{text; required}
Visit date
{date; required}
Team id
{integer; required}
Interviewer id
{integer; required}
Site/household-level data
Household occupied?
{char; required}
No (return later)/vacant
Yes
Consent given?
{char; required}
No
Yes
Study information
provided to respondent?
{char; required}
No
Yes
Household id
{alphanumeric (barcode); required}
double-entry
if manual entry
Respondent first-name,
{text; required}
Respondent last-name,
{text; required}
Respondent phone number,
{integer; required, 00 if not available}
double-entry
Respondent relation to head of household (HH)
{coded-text; required}
Respondent position in the household
{coded-text; required}
HH first-name
{text; required}
HH last-name
{text; required}
skip if respondent is
head of household
Household-level data
Member number
{integer; auto-incremental}
Member first-name
{text; required}
Member last-name
{text; required}
Member phone number
{integer; required, 00 if not available}
double
entry
Member ethnicity/tribe
{coded-text; required}
Member sex
{char; required}
Date of birth (DOB) or age?
{char; required}
Age
DOB
Age (month/year)
{integer; range}
DOB
{date; required}
Age/DOB < 16 years ?
{char; required}
Yes
No
relationship to HH
{coded-text; required}
Mother name
{popup-text; required}
Father name
{popup-text; required}
Attends school?
{char; required}
Repeat
if >1
relationship to HH
{coded-text; required}
Marital status
{coded-text; required}
Spouse name
{popup-text; required}
Level of education
{coded-text; required}
Employment status
{coded-text; required}
Individual-level data; {repeat until all members are enumerated; up to 20 members}
END


Malawi, Blantyre
Nepal, Kathmandu
Bangladesh, Dhaka
External storage
auto back-up
* MySQL database b_strataa
* MySQL database k_strataa
* MySQL database d_strataa
H
Auto-script: run nightly data sync
ODK Aggregate server
reports
P
Daily data uploads.
ODK Collect on android-based tablet
Auto-script: run daily anonymised data sync
Oxford University Vaccine Group (OVG) repository.


Bangladesh
Malawi
Nepal
% error
0.8
0.6
0.4
0.2
0.75
0.50
0.25
0.7
0.5
0.3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
fieldworker identification
period
Post-retraining
Pre-retraining


Post-retraining
Pre-retraining
Bangladesh
Malawi
Nepal
z-score
fieldworker identification
bad performance
good performance