

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

3

4 Authors: Deus Thindwa^{1,2}, Yama G Farooq³, Mila Shakya⁴, Nirod Saha⁵, Susan Tonks³, Yaw
5 Anokwa⁶, Melita A Gordon^{1,7}, Carl Hartung⁶, James E Meiring³, Andrew J Pollard³, Robert S
6 Heyderman^{1,8}, on behalf of The Strategic Typhoid alliance across Africa and Asia
7 consortium.

8

9 Affiliations

10 ¹ Malawi Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi.

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

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

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

16 ⁵ International Centre for Diarrhoeal Diseases Research, Dhaka, Bangladesh.

17 ⁶ Nafundi, Seattle, Washington, United States of America.

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

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

20

21 * Corresponding author

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

23

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

26

27 **Abstract**

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

33

34 **Results:** A census form was developed through a structured iterative process and
35 implemented using Open Data Kit Collect running on Android-based tablets. Data were
36 uploaded to Open Data Kit Aggregate, then auto-synced to MySQL-defined database nightly.
37 Data were backed-up daily from 3 sites centrally, and auto-reported weekly. Pre-census
38 materials' costs were estimated. Demographics of 308,348 individuals from 80,851
39 households were recorded within average of 14.7 weeks range (13-16) using 65 fieldworkers.
40 Overall, 21.7 errors (95% confidence interval: 21.4, 22.0) per 10,000 data points were found:
41 13.0 (95% confidence interval: 12.6, 13.5) and 24.5 (95% confidence interval: 24.1, 24.9)
42 errors on numeric and text fields respectively. These values meet standard quality threshold
43 of 50 errors per 10,000 data points. The EDC's total variable cost was estimated at
44 US\$13,791.82 per site.

45

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

48

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

50

51 **Background**

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

68

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

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

78

79 We present an open source-based EDC, designed to overcome data quality complexities, and
80 evaluate the efficiency, quality, and costs of the EDC by measuring volume, time, accuracy,
81 and material costs using multisite census data collected from sub-Saharan Africa and Asia.

82 The EDC was developed and implemented within the Strategic Typhoid alliance across
83 Africa and Asia (STRATAA), a comprehensive programme which is assessing population
84 dynamics and epidemiology of typhoid fever in Malawi, Bangladesh and Nepal to inform
85 design of vaccine and public health interventions.

86

87 **Design, development and implementation**

88 **Study setting and participants**

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

97 **Electronic census report form and data cleaning procedure**

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

111

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

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

146

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

148

149 **Electronic data capture system**

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

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

174

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

180

181 **Pre-census time, costs, and training**

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

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

191

192 **Statistical analysis and visualization**

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

212 worker, (μ_e) is the mean error for all fieldworkers, and (δ_e) is the error standard deviation.
213 In addition, we used Wilcoxon Signed-Rank Test for paired samples pre- versus post-
214 retraining in order to measure any statistical difference in the number of errors committed,
215 and determine whether retraining the fieldworkers helped improve accuracy. All statistics and
216 plots were conducted in R version 3.4.0 [21], eCRF flowchart and EDC diagram were created
217 using www.draw.io (JGraph, London, England).
218

219 **Results**

220 **Data collection volume, time and accuracy**

221 We recorded demographics of 308,348 individuals from 80,851 households in three countries
222 between June 2016 and October 2016; 97,410 individuals and 22,364 households from
223 Malawi, 100,207 and 32,368 from Nepal, and 110,731 and 26,119 from Bangladesh.
224 Completeness of household demographics enumeration were 94.2%, 75.6% and 79.2% for
225 Malawi, Nepal and Bangladesh, respectively, relative to background household count. The
226 average number of weeks for enumeration was 14.7 range (13-16) using 20, 25, 20 field
227 workers from Malawi, Nepal and Bangladesh, respectively. Overall, 21.7 errors (95%
228 confidence interval: 21.4, 22.0) per 10,000 data points were found; 15.9 errors (95%
229 confidence interval: 15.4, 16.4), 34.2 errors (95% confidence interval: 33.5, 34.9), and 14.6
230 errors (95% confidence interval: 14.2, 15.0) per 10,000 data points from Malawi, Nepal and
231 Bangladesh, respectively. Of the 17,707 errors documented from all sites, the majority 12,740
232 (72.0%) occurred on text fields compared to numeric fields 3,868 (21.8%). In addition, 1,099
233 (6.2%) errors occurred as duplicate records (e.g. either by enumerating a household or any of
234 its members at least twice) (Table 1).

235

236 Of all the data entry errors observed during enumeration period, 2,611 (65.4%), 6,265
 237 (65.8%) and 3,013 (71.8%) were, respectively, committed in Malawi, Nepal and Bangladesh
 238 prior fieldworkers' retraining. Moreover, there were fewer errors observed after retraining of
 239 fieldworkers compared to pre-retraining, and the differences were statistically significant in
 240 Malawi ($W = 5.5, P < 0.001$), Nepal ($W = 19.5, P < 0.001$), and Bangladesh ($W = 0, P < 0.001$)
 241 (Fig. 3).

242

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

244

245 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 [†]	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 [†]	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

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

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

253 □ Includes text, character, and date data types.

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

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

257

258 **Time and cost of census materials**

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

271

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

280

281 Table 2. Time and Costs Attainment Prior to Implementation of an Electronic Data Capture System
 282 (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	$X_1 \cdot Y$
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 †	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

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

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

285 § Some tablets already existed in other sites. Similarly, network devices and computer servers pre-existed in
 286 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

290 Discussion

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

297

298 Data were collected by largely secondary school level only fieldworkers receiving 1 week of
299 training and a day of retraining, and although the learning curve of using an eCRF in ODK
300 Collect on Android-based tablets was steep in the first 5 weeks of field work, high volume
301 and fairly accurate data were recorded (Fig. 3 and Fig. 4). The data accuracy of ~0.22%
302 errors (21.7 errors per 10,000 data points) reported in this study meets the acceptable quality
303 threshold of 50 errors per 10,000 data points recommended by the Society of Clinical Data
304 Management (SCDM, McLean, Virginia, USA) [22, 23]. The highly accurate EDC data in this
305 study is comparable to EDC data accuracies reported by the chronic disease research in South
306 Africa (0.17%) and maternal health survey in Burkina Faso (0.24%) [24, 25]. However, our
307 EDC data accuracy is superior to EDC data accuracies reported by the maternal health (2.8%)
308 and neglected tropic disease surveys (5.2%) in Ethiopia, the bloodstream infections study in
309 Zanzibar (1.0%) and the tuberculosis program in India (4.2%) [2, 3, 7, 26]. Moreover, our
310 EDC data are more accurate in comparison to data reported from paper-based studies of
311 maternal health (1.1%) and neglected tropical disease (6.2%) surveys in Ethiopia,
312 bloodstream infections study in Zanzibar (7.0%), chronic disease research in South Africa
313 (0.73%), and randomized controlled trial in Fiji (20.8%) [2–4, 7, 24]. As with previous
314 studies [2, 22, 27], text fields of this eCRF generated more errors than numeric fields, , and

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

320

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

322

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

330

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

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

345

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

359

360 Costs estimates on the data capture systems across low- and middle-income settings account
361 for different item inclusions [7, 28, 34, 35]. However, generally, our total variable cost of the
362 EDC was minimal relative to most EDCs or paper-based data collection methods conducted
363 in similar settings. For instance, our EDC's total variable cost is analogous to US\$13,883.00
364 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

380 **Availability and requirements**

381 The code scripts used to develop the EDC (ODK Collect eCRF and MySQL database
382 objects), and the raw data for errors analysed in this paper are all available through GitHub
383 [15].

384

385 **Abbreviations**

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

390

391 **Declarations**

392 **Ethics approval and consent to participate**

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

399

400 **Consent for publication**

401 Not applicable

402

403 **Availability of data and materials**

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

405

406 **Competing interests**

407 The authors declare that they have no competing interests.

408

409 **Funding**

410 Funding for the STRATAA study has been provided by a Wellcome Trust Strategic Award

411 (no. 106158/Z/14/Z), <https://wellcome.ac.uk/funding/managing-grant/grantsawarded>,

412 and the Bill and Melinda Gates Foundation (no. 617 OPP1141321),

413 <https://www.gatesfoundation.org/How-We-Work/Quick-Links/Grants-Database> to AJP. The

414 Malawi-Liverpool-Wellcome Programme and the Oxford University Clinical Research Unit

415 in Vietnam are supported by the Wellcome Trust with Major Overseas Programme core

416 awards. The funders did not play any role in the design of the study and collection, analysis

417 and interpretation of data and in writing the manuscript.

418

419 **Authors contribution**

420 Conceptualization: DT, YGF, RSH; Methodology: DT, YGF, MS, NS, YA, CH; Writing-

421 Original Draft: DT; Writing-Review and Editing: YGF, MS, NS, ST, YA, MAG, CH, JEM,

422 AJP, RSH; Funding-Acquisition: MAG, AJP, RSH. All authors read and approved the final

423 manuscript.

424

425 **Acknowledgements**

426 *Members of The Strategic Typhoid alliance across Africa & Asia consortium (STRATAA):*

427 Oxford University Clinical Research Unit, Patan Academy of Health Sciences, Kathmandu,

428 Nepal (Mila Shakya, Abhilasha Karkey, Sabina Dongol, Amit Aryjal, Buddha Basnyat);

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

451

452 **References**

- 453 1. Garritty C, Emam KE. Who's Using PDAs? Estimates of PDA Use by Health Care Providers: A
454 Systematic Review of Surveys. *J Med Internet Res.* 2006;8:e7.
- 455 2. Medhanyie AA, Spigt M, Yebyo H, Little A, Tadesse K, Dinant G-J, et al. Quality of routine health
456 data collected by health workers using smartphone at primary health care in Ethiopia. *Int J Med Inf.*
457 2017;101:9–14.
- 458 3. Thriemer K, Ley B, Ame SM, Puri MK, Hashim R, Chang NY, et al. Replacing paper data
459 collection forms with electronic data entry in the field: findings from a study of community-acquired
460 bloodstream infections in Pemba, Zanzibar. *BMC Res Notes.* 2012;5:113.
- 461 4. Yu P, de Courten M, Pan E, Galea G, Pryor J. The development and evaluation of a PDA-based
462 method for public health surveillance data collection in developing countries. *Int J Med Inf.*
463 2009;78:532–42.
- 464 5. Rorie DA, Flynn RWV, Grieve K, Doney A, Mackenzie I, MacDonald TM, et al. Electronic case
465 report forms and electronic data capture within clinical trials and pharmacoepidemiology. *Br J Clin*
466 *Pharmacol.* 83:1880–95.
- 467 6. Ali M, Deen JL, Khatib A, Enwere G, Seidlein L von, Reyburn R, et al. Paperless registration
468 during survey enumerations and large oral cholera mass vaccination in zanzibar, the United republic
469 of Tanzania. *Bull World Health Organ.* 2010;88:556–9.
- 470 7. King JD, Buolamwini J, Cromwell EA, Panfel A, Teferi T, Zerihun M, et al. A Novel Electronic
471 Data Collection System for Large-Scale Surveys of Neglected Tropical Diseases. *PLOS ONE.*
472 2013;8:e74570.
- 473 8. King C, Hall J, Banda M, Beard J, Bird J, Kazembe P, et al. Electronic data capture in a rural
474 African setting: evaluating experiences with different systems in Malawi. *Glob Health Action.*
475 2014;7:25878.
- 476 9. White A, Thomas DSK, Ezeanochie N, Bull S. Health Worker mHealth Utilization: A Systematic
477 Review. *Comput Inform Nurs CIN.* 2016;34:206–13.
- 478 10. Antillón M, Warren JL, Crawford FW, Weinberger DM, Kürüm E, Pak GD, et al. The burden of
479 typhoid fever in low- and middle-income countries: A meta-regression approach. *PLoS Negl Trop*
480 *Dis.* 2017;11:e0005376.
- 481 11. Mogasale V, Maskery B, Ochiai RL, Lee JS, Mogasale VV, Ramani E, et al. Burden of typhoid
482 fever in low-income and middle-income countries: a systematic, literature-based update with risk-
483 factor adjustment. *Lancet Glob Health.* 2014;2:e570–80.
- 484 12. Crump JA. Building the case for wider use of typhoid vaccines. *Vaccine.* 2015;33 Suppl 3:C1–2.
- 485 13. Darton TC, Meiring JE, Tonks S, Khan MA, Khanam F, Shakya M, et al. The STRATAA study
486 protocol: a programme to assess the burden of enteric fever in Bangladesh, Malawi and Nepal using
487 prospective population census, passive surveillance, serological studies and healthcare utilisation
488 surveys. *BMJ Open.* 2017;7:e016283.
- 489 14. MacPherson P, Choko AT, Webb EL, Thindwa D, Squire SB, Sambakunsi R, et al. Development
490 and Validation of a Global Positioning System–based “Map Book” System for Categorizing Cluster
491 Residency Status of Community Members Living in High-Density Urban Slums in Blantyre, Malawi.
492 *Am J Epidemiol.* 2013;177:1143–7.

- 493 15. STRATAA consortium. Electronic Data Capture for Large Scale Typhoid Surveillance. 2017.
494 [https://github.com/Oxfordvaccinegroup/Electronic-Data-Capture-for-Large-Scale-Typhoid-](https://github.com/Oxfordvaccinegroup/Electronic-Data-Capture-for-Large-Scale-Typhoid-Surveillance---STRATAA)
495 [Surveillance---STRATAA](https://github.com/Oxfordvaccinegroup/Electronic-Data-Capture-for-Large-Scale-Typhoid-Surveillance---STRATAA).
- 496 16. University of Washington. Open Data Kit. <https://opendatakit.org/>. Accessed 9 Jul 2017.
- 497 17. Anokwa Y, Hartung C, Brunette W, Borriello G, Lerer A. Open Source Data Collection in the
498 Developing World. *Computer*. 2009;42:97–9.
- 499 18. Hartung C, Lerer A, Anokwa Y, Tseng C, Brunette W, Borriello G. Open Data Kit: Tools to Build
500 Information Services for Developing Regions. In: Proceedings of the 4th ACM/IEEE International
501 Conference on Information and Communication Technologies and Development. New York, NY,
502 USA: ACM; 2010. p. 18:1–18:12. doi:10.1145/2369220.2369236.
- 503 19. Oracle Corporation, et al. MySQL. <https://www.mysql.com/>. Accessed 9 Jul 2017.
- 504 20. Chignell M, Tong T, Mizobuchi S, Delange T, Ho W, Walmsley W. Combining Multiple
505 Measures into a Single Figure of Merit. *Procedia Comput Sci*. 2015;69:36–43.
- 506 21. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for
507 Statistical Computing, Vienna, Austria. <https://www.r-project.org/>. Accessed 28 May 2019.
- 508 22. Jenkins TM, Boyce TW, Akers R, Andringa J, Liu Y, Miller R, et al. Evaluation of a Teleform-
509 based data collection system: A multi-center obesity research case study. *Comput Biol Med*.
510 2014;49:15–8.
- 511 23. Pomerantseva V, Ilicheva O. Clinical Data Collection, Cleaning and Verification in Anticipation
512 of Database Lock. *Pharm Med*. 2011;25:223–33.
- 513 24. Dillon DG, Pirie F, Rice S, Pomilla C, Sandhu MS, Motala AA, et al. Open-source electronic data
514 capture system offered increased accuracy and cost-effectiveness compared with paper methods in
515 Africa. *J Clin Epidemiol*. 2014;67:1358–63.
- 516 25. Byass P, Hounton S, Ouédraogo M, Somé H, Diallo I, Fottrell E, et al. Direct data capture using
517 hand-held computers in rural Burkina Faso: experiences, benefits and lessons learnt. *Trop Med Int*
518 *Health*. 2008;13:25–30.
- 519 26. Patnaik S, Brunskill E, Thies W. Evaluating the Accuracy of Data Collection on Mobile Phones:
520 A Study of Forms, SMS, and Voice. 2009. <https://www.cs.cmu.edu/~ebrun/patnaik-ictd09.pdf>.
521 Accessed 22 May 2017.
- 522 27. Walther B, Hossin S, Townend J, Abernethy N, Parker D, Jeffries D. Comparison of Electronic
523 Data Capture (EDC) with the Standard Data Capture Method for Clinical Trial Data. *PLOS ONE*.
524 2011;6:e25348.
- 525 28. McLean E, Dube A, Saul J, Branson K, Luhanga M, Mwiba O, et al. Implementing electronic data
526 capture at a well-established health and demographic surveillance site in rural northern Malawi. *Glob*
527 *Health Action*. 2017;10:1367162.
- 528 29. Rajput ZA, Mbugua S, Amadi D, Chepneno V, Saleem JJ, Anokwa Y, et al. Evaluation of an
529 Android-based mHealth system for population surveillance in developing countries. *J Am Med*
530 *Inform Assoc JAMIA*. 2012;19:655–9.

- 531 30. Maduka O, Akpan G, Maleghemi S. Using Android and Open Data Kit Technology in Data
532 Management for Research in Resource-Limited Settings in the Niger Delta Region of Nigeria: Cross-
533 Sectional Household Survey. *JMIR MHealth UHealth*. 2017;5. doi:10.2196/mhealth.7827.
- 534 31. Meyer J, Fredrich D, Piegsa J, Habes M, van den Berg N, Hoffmann W. A mobile and
535 asynchronous electronic data capture system for epidemiologic studies. *Comput Methods Programs*
536 *Biomed*. 2013;110:369–79.
- 537 32. Baguiya A. An offline mobile data capture module for health and demographic surveillance
538 system (HDSS) studies. Thesis. 2016. <http://wiredspace.wits.ac.za/handle/10539/21399>. Accessed 18
539 Jul 2017.
- 540 33. Shirima K, Mukasa O, Schellenberg JA, Manzi F, John D, Mushi A, et al. The use of personal
541 digital assistants for data entry at the point of collection in a large household survey in southern
542 Tanzania. *Emerg Themes Epidemiol*. 2007;4:5.
- 543 34. Flaxman AD, Stewart A, Joseph JC, Alam N, Alam SS, Chowdhury H, et al. Collecting verbal
544 autopsies: improving and streamlining data collection processes using electronic tablets. *Popul Health*
545 *Metr*. 2018;16:3.
- 546 35. Njuguna HN, Caselton DL, Arunga GO, Emukule GO, Kinyanjui DK, Kalani RM, et al. A
547 comparison of smartphones to paper-based questionnaires for routine influenza sentinel surveillance,
548 Kenya, 2011–2012. *BMC Med Inform Decis Mak*. 2014;14:107.
- 549







