**DClare cluster RCT open access data: Information and instructions for use**

*Description of the data:*

The Bangladesh Diabetes: Community-Led Awareness, Response and Evaluation (DClare) project was a collaboration between University College London Institute for Global Health, the Diabetic Association of Bangladesh and Karolinska Institutet’s Department of Global Public Health. The DClare project involved a cluster randomised controlled trial (cRCT), which was conducted in Alfadanga Upazilla, Faridpur District, Bangladesh between 2019 and 2022. This document describes the survey data from this cRCT which has been made openly accessible.

Other datasets were collected for the DClare Project which have not been openly archived. Please refer to the individual publications for information on how to access these data.

* Stakeholder perspectives on intervention scale-up: [10.1186/s12889-023-15551-9](https://doi.org/10.1186/s12889-023-15551-9)
* Community engagement in the context of COVID-19: <https://doi.org/10.14324/RFA.07.1.17>
* Process evaluation of intervention implementation during COVID-19: <https://doi.org/10.1136/bmjopen-2024-089288>

The full methods for the cRCT have been described in the published study protocol, trial registration and main trial results publications:

* Original trial protocol for a stepped-wedge trial: <https://doi.org/10.1186/s13063-021-05167-y>
* Updated trial protocol for a cRCT, due to the COVID-19 pandemic: <https://doi.org/10.1186/s13063-023-07243-x>
* Trial registration: <https://doi.org/10.1186/ISRCTN42219712>
* Main trial findings: To be added

Briefly, Alfadanga upazila was divided into 12 clusters of approximately equal population size using government census data. Within these 12 clusters, between 2-5 villages were purposefully selected to achieve between 800-1000 households per cluster, according the following eligibility criteria: a) they do not sit on a border with a neighbouring study cluster; b) they are not a major trading centre or administrative centre; c) they have a minimum of 50 households. A population census was conducted between November 2019 – January 2020 to form a sampling frame, with each household within purposefully selected villages visited, and all household members aged 25 and older registered. A unique study ID was assigned to each individual using the following formulation: *Cluster ID (2 digits) + village ID (2 digits) + household number (3 digits) + person number (1 digit).*

Three cross-sectional surveys were conducted: a) pre-COVID-19 baseline; b) post-COVID-19 baseline; c) endline. For each survey, a random sample of individuals were selected using two-stage simple random sampling, with the household randomly selected then one eligible individual selected from these households. For the endline survey, individuals who were identified as living with intermediate hyperglycaemia in the post-COVID-19 baseline survey were purposefully selected. **Therefore, individuals with intermediate hyperglycaemia could be sampled twice for the endline – randomly and purposefully. The random sample was re-done for each survey; therefore individuals could be sampled for more than one of these surveys by chance.** The table below indicates the sample for each survey, and the completion rates.

|  |  |  |
| --- | --- | --- |
| **Survey** | **Sampled individuals** | **Completed recruitment of eligible participants** |
| a) Old baseline (February – March 2020) | Total: 1320 | Anthropometry: 948  Questionnaire: 948  Total: 950 |
| b) New baseline (January – March 2021) | Total: 1584 | Anthropometry: 1392  Questionnaire: 1392  Total: 1392 |
| c) Endline (September – November 2022) | Random: 1584  Intermediate hyperglycaemia cohort: 313  Total: 1868 | Anthropometry: 1589  Questionnaire: 1589  Total: 1589 |

Each cross-sectional survey round involved two surveys: a) anthropometry; b) questionnaire. Individuals were consented to each survey separately and were allowed to take part in one without completing the other. Information about the content of each survey and how the data were collected are described in the protocol, study registration and main trial paper.

*List of files:*

Three data files, in both .csv and Stata .dta formats, and meta-data are available:

* DClareOldBaseline.csv, DClareOldBaseline.dta
* DClareNewBaseline.csv, DClareNewBaseline.dta
* DClareEndline.csv, DClareEndline.dta
* OldBaselineQuestionnaireMetadata.xlsx
* NewBaselineQuestionnaireMetadata.xlsx
* EndlineQuestionnaireMetadata.xlsx
* AnthropometryMetadata.xlsx

*Using the data:*

Unique ID:

The original study IDs have been removed from these datasets, to ensure the open archive data are fully anonymous, along with any information about individual names, address, contact and GPS. Village name has been replaced with a sequential code (‘res\_village\_anon’) and a new individual ID number (‘res\_id\_anon’) was generated using a random number calculator within Stata, creating a unique 6-digit code that cannot be linked to the individuals cluster, village, or household.

Personal information that is still included for each individual includes: cluster name, age in years, self-reported gender and self-reported clinical data. We deemed these data points insufficient to identify an individual.

Weights:

Due to the sampling approach, for analyses that want to present population estimates, rather than sample estimates, weights should be used. The weights differ for each survey, given the sample size changed and the eligible population changed. The weights should only be used for analyses which use the randomly sampled population – analyses of the purposefully sampled cohort of individuals identified with intermediate hyperglycaemia should not. Two types of weights are provided: a) sample weights (‘sample\_wt’), i.e. the inverse probability of an individual being sampled from that cluster; b) standardized weights (‘standard\_wt’), i.e. the population adjusted weight.

Sample:

There are two variables in each dataset which identify whether the individual was sampled through the simple random sampling approach (‘ran\_sample’), or purposeful sampling for those with intermediate hyperglycaemia (‘res\_pred’). A value of 1 indicates that the participant was selected using this method. All individuals will have at least one of these selected, and some individuals will have both selected.

Linking surveys:

The three cross-sectional surveys can be merged, to create a longitudinal analysis dataset. As some individuals were sampled more than once (either purposefully for intermediate hyperglycaemia, or randomly by chance), the unique anonymised individual ID should be used, with a 1:1 matching. Note when merging datasets, instead of appending them, that values within the same variables are expected to differ, and so renaming variables with a suffix to indicate which survey they come from is recommended (e.g. renaming the first baseline survey variables to be “old\_X”).

Acknowledgement:

The DClare trial was funded by the Medical Research Council UK (MR/T023562/1) under the Global Alliance for Chronic Diseases Scale-Up Programme.

The DClare trial team consisted of: Edward Fottrell, Abdul Kuddus MPH, Joanna Morrison, Tasmin Nahar, Carina King, Sanjit Kumer Shaha, Malini Pires, Sarker Ashraf Uddin Ahmed, James Beard, Naveed Ahmed, Andrew Copas, Hassan Haghparast-Bidgoli, A.K. Azad Khan, and Kishwar Azad

We are grateful to Md Golam Azam and all data collectors and supervisors for their contributions to the surveys that generated these data.

Any publications which use these data should include the following acknowledgement statement: *“Data used in this study were made available as part of the Bangladesh DClare project – a collaboration between University College London Institute for Global Health, the Diabetic Association of Bangladesh Centre for Health Research and Implementation, and Karolinska Institutet’s Department of Global Public Health, funded by the Medical Research Council UK (MR/T023562/1) under the Global Alliance for Chronic Diseases Scale-Up Programme”.*

For further information about the data which cannot be found within this document or published papers, please contact:

* Prof Ed Fottrell: [e.fottrell@ucl.ac.uk](mailto:e.fottrell@ucl.ac.uk)
* Prof Kishwar Azad: [kishwar.azad@googlemail.com](mailto:kishwar.azad@googlemail.com)
* A/Prof Carina King: [carina.king@ki.se](mailto:carina.king@ki.se)