Check for updates

DATA NOTE

REVISED Family network and household composition: a longitudinal dataset derived from the Karonga Health and Demographic Surveillance System, in rural Malawi [version 2; peer review: 3 approved]

Estelle McLean[1,2], Fredrick Kalobekamo[2], Oddie Mwiba[2], Amelia C Crampin[1,2], Emma Slaymaker[1], Rebecca Sear[1], Albert Dube[2]

[1]Department of Population Health, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK
[2]Malawi Epidemiology and Intervention Research Unit, Chilumba, Karonga District, Malawi

v2 **First published:** 14 Dec 2023, **8**:573
https://doi.org/10.12688/wellcomeopenres.20406.1
Latest published: 30 Apr 2024, **8**:573
https://doi.org/10.12688/wellcomeopenres.20406.2

Abstract

Proximity to family, household composition, and structure are often studied as outcomes and as explanatory factors in a wide range of scientific disciplines. Here, we describe a large longitudinal dataset (currently including data from over 70,000 individuals from 2004 to 2017), including data on household structure, proximity to kin, population density, and other socio-demographic factors derived from data from the Karonga Health and Demographic Surveillance Site (HDSS) in Northern Malawi. We present how the dataset is generated, list some examples of how it can be used, and provide information on the limitations that affect the types of analyses that can be carried out.

Keywords

Family, Relatives, Household, GPS, Longitudinal, Malawi

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 2 (revision) 30 Apr 2024	✓ view	✓ view	✓ view
version 1 14 Dec 2023	? view		

1. **Ashira Menashe-Oren**, Universite catholique de Louvain, Louvain-la-Neuve, Belgium
2. **Laurie DeRose**, The Catholic University of America, Washington, USA
3. **Sarah Reynolds**, University of California Berkeley, Berkeley, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Estelle McLean (estelle.mclean@lshtm.ac.uk)

Author roles: McLean E: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kalobekamo F**: Investigation, Project Administration, Resources; **Mwiba O**: Data Curation; **Crampin AC**: Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; **Slaymaker E**: Supervision, Writing – Review & Editing; **Sear R**: Conceptualization, Supervision, Writing – Review & Editing; **Dube A**: Data Curation, Investigation, Project Administration, Resources, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome (grants 098610, a strategic award to Moffat Nyirenda and the Karonga HDSS and 217073 a Longitudinal Population Study grant awarded to Amelia Crampin and the Karonga HDSS).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: McLean E, Kalobekamo F, Mwiba O *et al.* **Family network and household composition: a longitudinal dataset derived from the Karonga Health and Demographic Surveillance System, in rural Malawi [version 2; peer review: 3 approved]** Wellcome Open Research 2024, **8**:573 https://doi.org/10.12688/wellcomeopenres.20406.2

First published: 14 Dec 2023, **8**:573 https://doi.org/10.12688/wellcomeopenres.20406.1

REVISED **Amendments from Version 1**

A table of summaries by calendar year for selected variables has been added, to help readers understand the dataset (Table 1). A link to the documentation of the dataset with the full list and description of all variables has also been added. Minor edits have been made to the "Relationships within households" section to try to make this clearer, a diagram has also been added to improve this section (Figure 1). A table of the distribution of participants in the dataset by birth cohort and time resident in the HDSS has been converted to a population pyramid (Figure 4). A summary paragraph has been added, and minor changes to the text have been made to clarify some methodological decisions.

Any further responses from the reviewers can be found at the end of the article

Introduction

Proximity to family, household composition, and structure have been studied and described as outcomes themselves (Keilman, 1988) and as explanatory factors in a diverse range of disciplines, including nutrition (Bronte-Tinkew & Dejong, 2004), childhood vaccination (Gage *et al.*, 1997), poverty (Snyder *et al.*, 2006), education (Perkins, 2019), evolutionary biology (Flinn *et al.*, 2007), criminology (Maxfield, 1987), child abuse (Stiffman *et al.*, 2002), transportation (Strathman *et al.*, 1994) and tourism (Tangeland & Aas, 2011). This data note describes a large longitudinal dataset (currently including data from over 70,000 individuals from 2004 to 2017), including data on household structure, proximity to kin, population density, and other socio-demographic factors derived from data from the Karonga Health and Demographic Surveillance Site (HDSS) in Northern Malawi. The Karonga HDSS is run by the Malawi Epidemiology and Intervention Research Unit (MEIRU), formerly known as the Karonga Prevention Study. It was set up between 2002 and 2004, but built upon research infrastructure that has been ongoing in the same area since 1979 (Ponnighaus *et al.*, 1987). Early research in the area focused on leprosy, and as the disease was known to cluster in families, considerable effort was expended on linking research participants (with and without leprosy) to their parents to generate family lineages. This practice has continued to the present day, allowing the generation of this rich dataset.

Methods

Context

The Karonga HDSS was established between 2002 and 2004 in the southern Karonga district of northern Malawi (Crampin *et al.*, 2012). The area is largely rural, with one semi-urban trading town, several smaller market villages, and one port in Lake Malawi. The majority of the population engage in subsistence farming or fishing. The main ethnic group living here is Tumbuka, who since the 19th century have followed patrilineal and patrilocal customs: women tend to move to their husbands' villages when they marry (Malawi Human Rights Commission, 2006). In the event of divorce or even paternal death, children who are old enough to be away from their mothers may be required to live with their fathers' families (Malawi Human Rights Commission, 2006). Polygyny is widespread; at the end of 2016, about 15% of households in the HDSS were headed by men with more than one wife.

Initial data

The HDSS covers an area of 150km^2 and by 2016, over 40,000 people were under surveillance. Births and deaths are captured monthly through a system of local 'key informants,' whereas migrations are captured annually through visits to all households. Specific dates for each event are captured; therefore, the data are arranged as episodes that may start with the initial census, birth, or in-migration and end with death or out-migration. Participants are given a unique identifier (ID) that they retain in all studies: if they move, they are linked back to this ID (even if they left the area and then returned). Households are also given unique identifiers and the household ID is listed as part of each residency episode. If a participant moves to a new household within the area, the episode at the old household ends, and a new one begins. In the HDSS, a household is defined as a group of individuals, rather than a location, meaning that if the group moves, they would still be classified as the same household. Household membership is defined by the participants under the guidance of trained fieldworkers: all household members must usually live in the dwelling/compound together and recognize the same household head (Crampin *et al.*, 2012). Men with more than one wife who do not live in the same location are assigned to live in all the co-wives' households; all other participants may only belong to one household. GPS coordinates were recorded for each household at the initial census, when the household is established, and if it moves. House move or change in household membership may result in one household being 'dissolved' and other(s) established. Because the household ID is listed with each person's residency episode, it is possible to link all individual household members at any time point.

When a new HDSS participant is registered through birth or in-migration, where possible, members of any age are linked to their parents' identification numbers if they have ever been assigned one. On an annual basis, participants are asked about their marital status and to provide information about their spouse(s), where possible, the identification numbers of the spouses have also been linked. Parents and spouses do not need to be HDSS members themselves to receive an identifying number.

Regular and one-off surveys have been carried out in the area using the HDSS as a platform. Individual and household socio-economic status variables are regularly gathered.

Data processing

Raw data are currently stored in Microsoft Access databases and extracted in the Stata format. All data processing to create this dataset described in this paper was performed using Stata 16.1.

The longitudinal dataset described in this paper is in the format of an unbalanced panel dataset, with HDSS residents contributing one record for each period while they were living in the HDSS area from 2004 (the first complete year of complete surveillance) to 2017. The residency episodes are first reduced to one record per person per period by taking a snapshot at the midpoint of the period. This allows for more flexible data manipulation: the rate of change of the time-varying variables which are added to this dataset can be extremely high so maintaining the data in episodic format would make the data manipulations very complex and too computationally intense. As continuous data are available for all HDSS residents, the length of the period represented by the snapshot can vary according to the needs of the analysis (*i.e.*, yearly, quarterly, or monthly). This description uses a mid-year snapshot as an example, but the same processes can be used for any period.

Separately, the parent ID and spouse-ID lists are combined to generate a long list of all blood and non-blood relationships between all HDSS residents. Each relationship record includes the detailed relation type (e.g., mother, half-sister, great-aunt), family type: maternal (mother and any relatives through her [*i.e.* grandparents, aunts/uncles and cousins), paternal (father and any relatives through him), sister (half or full sister and any of her children or grandchildren), brother (half or full brother and any of his children or grandchildren), daughter (daughter and any of her children or grandchildren), son (son and any of his children or grandchildren), the estimated genetic relatedness (*i.e.*, 50% for parent-child, down to 3.125% for mother's cousin), categorized age difference and sex of the relation. For blood relationships, the most distant included were children of cousins and mother's or father's cousin; for non-blood relatives, step-family was included up to step-great grandparent/child (though not other step-relations *i.e.,* step-cousins or aunts), spouse, spouse's family (in-laws), and spouses of blood relatives, both up to cousins/great-grandparents. Being related in more than one way is possible in this area; for example, a widow may marry her deceased husband's brother, so for her children from the first marriage, the new husband would be both their uncle and step-father. One 'closest' relationship was selected as the main one by preferring blood over non-blood relationships and, within the blood relatives, choosing the one with the highest average genetic relatedness. The full list of relations for people with more than one link is also available.

The population panel and the relationships dataset are used in three linked processes that generate variables describing household characteristics, the relationships between the index person and their other household members, and their family network beyond the household. The resulting datasets from the three processes are merged so that all of the above information is available for each person, at each time point that they are present in the HDSS.

Household characteristics

The population panel data were used to create a summary dataset describing the households at each time point. All households in each mid-year snapshot were first summarized into the number of household members by age group. The age composition of households can be used as an indicator of vulnerability, *i.e.* by calculating the number of working-age adults to dependent children and older adults. Second, the average relatedness between all household members is calculated, which is a measure of kinship within social groups often used in social biology (Koster, 2018). Finally, the proportion of all the relationships in the household that are unknown is calculated, which is when there is no known blood or non-blood relationship between them, but either one lacks at least one parental ID, so we cannot be sure that they are non-relatives. This is an indicator of the data quality.

The distance between each index household and every other household in the local area is then calculated. The summary household variables were then used to calculate, for each household at each time point, the number of other households, the number of other people (overall and by age group), the mean household relatedness, and the mean level of missing data within certain radii (*i.e.* 25m, or 250m). These are indicators of population density, several different radiuses are used to reflect the types of habitation that the HDSS covers, to be able to differentiate between households living in the dense trading centre (high density in both narrow and wider radius), in small, isolated clusters of households (high density in narrow radius, but low in wider radius) or in loosely connected villages (medium density in both narrow and wide radius). The population density variables were also used to identify linked households in the analyses (see below).

Relationships within households

While people in Karonga mostly do not live in shared compounds, as is common in other settings, it was known from field worker reports and through interrogation of the data that two or more households sometimes reside in very close proximity, sharing facilities in loose economic or social alliances, with shared resources and linked prospects. Using the population panel and relationship data, these grouped households were identified to generate an 'expanded household' definition, in addition to the standard household definition as used in HDSS operations (referred to as the 'immediate household' in this paper). Grouped households were not formally identified during surveillance; thus, a data-driven approach was used to harness the spacing between households at different population densities together with relationship data.

To start to develop an algorithm to identify linked households, a random sample of 100 pairs of households 30m or less (but over 0m) apart was examined individually using satellite imagery on Google Earth and assigned by eye as being in the same or in different compounds. This exercise showed that the 'same' compound households were a median of 7.7m (range 1.7–21.2m, IQR 4.1–11.4) apart while the 'different' compound households were 18m (range 6–29.5m, IQR 13.7–21.3) apart. Thus, it was assumed that, across the HDSS in the full dataset, all households less than 5m apart were linked and may be linked if they were up to 20m apart. The likelihood of households between 5 and 20 metres apart being in the same compound depended on the density of households in the area: i.e. 2 households 10m apart in a sparsely population area were very likely to be in the same compound, however in a densely populated area this likelihood is very low. Thus, an iterative algorithm was created which initially assigned households a 'guide radius' of 5, 10, 15 or 20 metres, according to how many other households were present within 50m, and whether the number of households near to them was as expected assuming an even distribution of the households within the 50m radius. For example, a household with 20 households within 50m (7852 m^2) would expect to have 0.8 households within 10m (314.2 m^2) and 3.2 within 20m (1256.6 m^2) if they are distributed equally, so if they have two households within 10m and three within 20m the initial radius would be set at 10m. If a household had the expected number of households according to the 50m radius it was given a starting radius of 5m. Households within the guide radius were linked if there was at least one relationship link between the households (*i.e*., at least one member of one household is related by blood or marriage to at least one member of the other household) (Figure 1). This method is prone to error but results in more appropriate connections between households than using a simpler rule such as all households within 5m (which would reduce the number of connections made in more rural areas where linked buildings can be more spaced out) or within 20m (which would inappropriately connect multiple households in more densely populated areas).

Once all members of each individual's 'immediate' (as recorded in the data) and 'expanded' (as described above) households were identified, the listing of all blood and non-blood relationships was used to create binary or continuous variables indicating the presence of certain relative types, *i.e*. mother in immediate household, or number of maternal half siblings aged under 18 in expanded households.

Family network

The GPS coordinates of all blood relatives (either singly or as groups, *i.e*., maternal or paternal) were compared to those of the index at each time point. Summary variables were then calculated as either binary or continuous for the presence of relatives within certain radii (e.g., father living with 250m, number of maternal aunts aged over 18 living within 100m, number of paternal relatives living within 50m). These variables are named and coded similarly to household-relative variables.

Examples of uses of dataset

The full listing of variables in the dataset may be found on the MEIRU data catalogue (https://kpsmw.lshtm.ac.uk/index.php/catalog/13). Summaries of some example variables from the dataset over calendar year are shown in Table 1. This dataset has been used in an in-depth analysis of household composition, including an assessment of whether latent class analysis can be used to create data-driven household classifications (McLean *et al.*, 2021a), an analysis of transition to adulthood by using household composition variables to identify when an adolescent can be described as having left home (along with other variables related to leaving school, getting married, and having children) in a sequence analysis (McLean *et al.*, 2021b) and an analysis of the effect of the presence of family within and outside the household on short and long migration in children and adolescents (McLean *et al.*, 2023). Other analyses related to mortality and fertility are possible, and as the HDSS is ongoing, more analyses linking childhood household composition/structure with adult outcomes will be possible. Newly collected data can be added to the datasets by re-running all the processes with the updated datasets. Other HDSSs collect similar data, and thus may be able to generate similar datasets, following the logic described above.

Dataset validation / limitations

Although this dataset has many potential uses, it is important for users to be aware of some limitations to aid in the appropriate selection of data for analyses. The dataset is dependent on parent and spouse links, which are not available

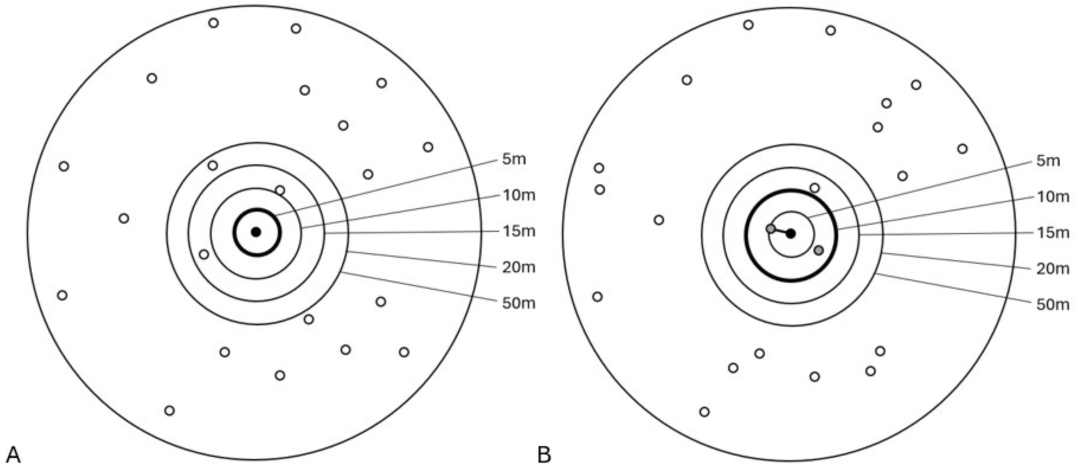

Figure 1. Example of household linkage: in both **A** & **B** there are 20 households within a 50m radius of the black index household; in **A** the households are evenly spread and the number within the smaller radii are as expected, the guide radius of 5m is assigned and there are no linked households; In **B** there are more households within 10m than would be expected so the guide radius of 10m is assigned, relationships between the members of the black index and the grey potentially linked households are checked, and a link made with one of them.

Table 1. Summaries of selected variables from the dataset by year.

	2004	2008	2012	2016
Household indicators				
Mean household size	5.1	4.8	4.6	4.5
Median household size	5	5	4	4
Mean genetic relatedness within household	19.1%	19.3%	19.4%	19.1%
Mean number of births in household	0.23	0.19	0.16	0.13
Mean number of deaths in household	0.05	0.03	0.03	0.02
Individual indicators (age indicated in brackets)				
% living with maternal grandmother (u15y)	12.0%	10.4%	10.2%	10.7%
% living within 50m (but not same household) of maternal grandmother (u15y)	0.5%	0.8%	1.1%	1.5%
% living with child aged 18 or over (60y+)	46.7%	40.5%	38.2%	38.9%
% living within 50m (but not same household) of child aged 18 or over (60y+)	17.9%	22.2%	24.9%	25.1%
Mean number of maternal relatives living with 250m (15–29y)	1.0	1.1	1.1	1.2
Mean number of Paternal relatives living with 250m (15–29y)	2.5	2.7	2.9	3.3
Mean genetic relatedness of household members to index (u15y)	36.1%	37.8%	39.4%	38.8%
Mean genetic relatedness of household members to index (15–29y)	25.9%	27.0%	28.5%	29.2%
Mean genetic relatedness of household members to index (30–44y)	31.6%	33.1%	33.5%	33.7%
Mean genetic relatedness of household members to index (45–59y)	29.2%	29.3%	29.9%	30.8%
Mean genetic relatedness of household members to index (60y+)	19.2%	18.9%	19.3%	20.3%
% of people living within 250m who are blood relatives (all ages)	13.5%	13.0%	12.6%	11.7%

for all HDSS members. The proportion of all HDSS members by age and whether their mother and father IDs are known is shown in Figure 2. Children had the highest proportion of known parental IDS, and there was very good coverage for the youngest children. After childhood, the proportion with no IDs is relatively stable at around 30%, with most people having both mother and father ID available.

Being able to link individuals to their relatives also depends on whether other people have parent/spouse ID links. Figure 3 shows the average proportion of household relationships unknown (due to missing IDs meaning that we cannot confidently assign the pair as unrelated) by the age of the index person and calendar year. Unsurprisingly, the group with the lowest proportion of unknown relationships is children aged under five, but the 30–49-year age group also has low levels (as their households are likely to be formed of their spouse and children). The groups with the highest proportion of unknown relationships were people aged over 70 years and adolescents aged 15–19, however the proportions were not high (under 13%). By calendar year, the proportions unknown decreased somewhat from 2004, but there was an increase at the end of the period to 2017.

The actual number of individuals available in the dataset by year and age group is shown in Table 2, which also shows the proportion with complete information on their relationships with all household members and the proportion with no information at all. This shows that there are a high number of individuals with sufficient data in all age groups, although the numbers decrease after the age of 50 years.

Another potential limitation of the dataset is related to the HDSS data source: data are only available on participants when they live in the HDSS area. Figure 4 shows the number of HDSS residents by sex and birth cohort and the number of years they were present in the HDSS between 2004 and 2017 (maximum 14 years). While a large number of participants have complete data for the whole 14-year period, it is important to note that those who remain in the area are likely to be different from those who do not. Figure 4 shows the effects of birth cohort (those with earlier birth dates are more

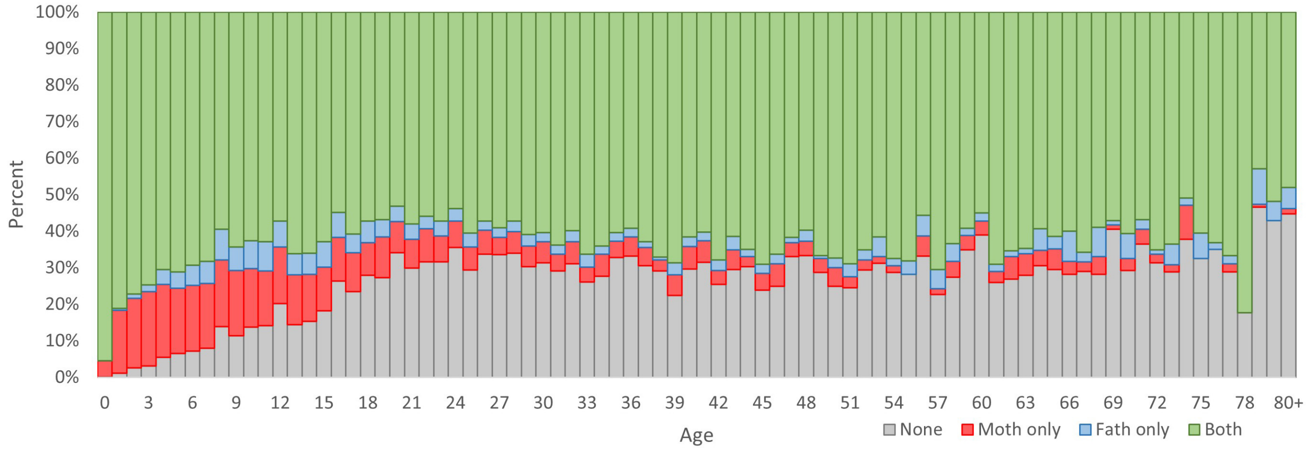

Figure 2. Percent of HDSS residents by age and availability of parent ID-links. Percent of participants with no parent ID-links are at the bottom of the columns in grey, above that in red are those with only mother ID-link, above that in blue are those with only father ID-link, and at the top in green are those with both parent ID-links.

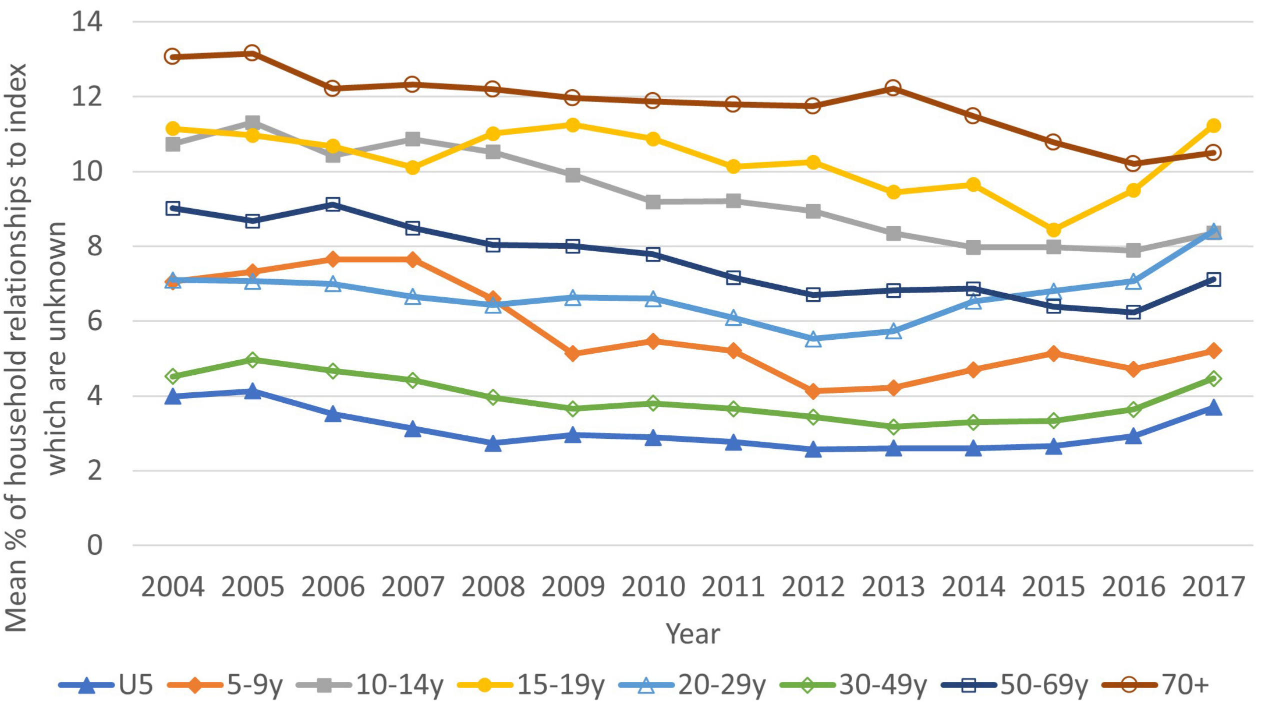

Figure 3. Average percent of relationships to index person within households which are unknown, by age group (of index) and year. Data for under-fives is shown with solid blue triangles, for five-nine year-olds with solid orange diamonds, for 10–14 year-olds with solid grey squares, for 15–19 year-olds with empty yellow circles, for 20–29 year-olds with empty blue triangles, for 30–39 year-olds with empty green diamonds, for 50–59 year-olds with empty dark blue squares and for 70+ with empty maroon circles.

likely to have complete data) and sex (males are more likely to have complete data).

Conclusion

This complex dataset allows for many analyses of family and household structure and kin proximity in rural northern Malawi. Linkages within the HDSS also mean that further variables may be available to link to this dataset, at certain time points. The HDSS is ongoing, so the dataset may also be updated with more recent data when possible, and the dataset could be used as a sampling frame to identify participants for further primary data capture.

Table 2. Total number of individuals* in the dataset by age group, selected years and whether their relationship to other household members are fully known or fully unknown.

Age group	Households		2005	2007	2009	2011	2013	2015	2017
All	Total	n	31596	33685	34027	35833	37787	40453	43523
	Fully known	n	25112	27343	28250	30169	32272	34334	35864
		%	79.5%	81.2%	83.0%	84.2%	85.4%	84.9%	82.4%
	Fully unknown	n	1017	1062	1025	1019	1025	1134	1488
		%	3.2%	3.2%	3.0%	2.8%	2.7%	2.8%	3.4%
U5	Total	n	6161	6671	6437	6350	6265	6409	6311
	Fully known	n	5281	5886	5762	5738	5729	5852	5607
		%	85.7%	88.2%	89.5%	90.4%	91.4%	91.3%	88.8%
	Fully unknown	n	60	37	42	34	43	42	64
		%	1.0%	0.6%	0.7%	0.5%	0.7%	0.7%	1.0%
5–9y	Total	n	4696	5424	5315	6033	6230	6473	6565
	Fully known	n	3831	4421	4554	5217	5512	5661	5659
		%	81.6%	81.5%	85.7%	86.5%	88.5%	87.5%	86.2%
	Fully unknown	n	188	231	128	159	128	183	168
		%	4.0%	4.3%	2.4%	2.6%	2.1%	2.8%	2.6%
10–14y	Total	n	4107	4297	4418	4847	4979	5852	6631
	Fully known	n	3121	3326	3520	3959	4138	4880	5442
		%	76.0%	77.4%	79.7%	81.7%	83.1%	83.4%	82.1%
	Fully unknown	n	306	317	298	300	281	310	354
		%	7.5%	7.4%	6.7%	6.2%	5.6%	5.3%	5.3%
15–19y	Total	n	2943	2883	3473	3435	4115	4288	4818
	Fully known	n	2208	2241	2730	2740	3389	3536	3757
		%	75.0%	77.7%	78.6%	79.8%	82.4%	82.5%	78.0%
	Fully unknown	n	187	177	268	234	277	241	371
		%	6.4%	6.1%	7.7%	6.8%	6.7%	5.6%	7.7%
20–29y	Total	n	5315	5775	5076	5529	5537	6266	6889
	Fully known	n	4261	4733	4209	4687	4753	5250	5560
		%	80.2%	82.0%	82.9%	84.8%	85.8%	83.8%	80.7%
	Fully unknown	n	145	176	159	174	170	237	329
		%	2.7%	3.0%	3.1%	3.1%	3.1%	3.8%	4.8%
30–49y	Total	n	5202	5514	5943	6273	6966	7340	8045
	Fully known	n	4239	4568	5084	5389	6044	6336	6746
		%	81.5%	82.8%	85.5%	85.9%	86.8%	86.3%	83.9%
	Fully unknown	n	55	49	44	48	40	48	106
		%	1.1%	0.9%	0.7%	0.8%	0.6%	0.7%	1.3%

Age group	Households		2005	2007	2009	2011	2013	2015	2017
50–69y	Total	n	2173	2128	2328	2376	2605	2805	3038
	Fully known	n	1558	1527	1713	1799	1999	2151	2288
		%	71.7%	71.8%	73.6%	75.7%	76.7%	76.7%	75.3%
	Fully unknown	n	42	36	46	35	42	38	57
		%	1.9%	1.7%	2.0%	1.5%	1.6%	1.4%	1.9%
70+	Total	n	999	993	1037	990	1090	1020	1226
	Fully known	n	613	641	678	640	708	668	805
		%	61.4%	64.6%	65.4%	64.6%	65.0%	65.5%	65.7%
	Fully unknown	n	34	39	40	35	44	35	39
		%	3.4%	3.9%	3.9%	3.5%	4.0%	3.4%	3.2%

Note that individuals contribute data to this table for all years which they are present in the HDSS.

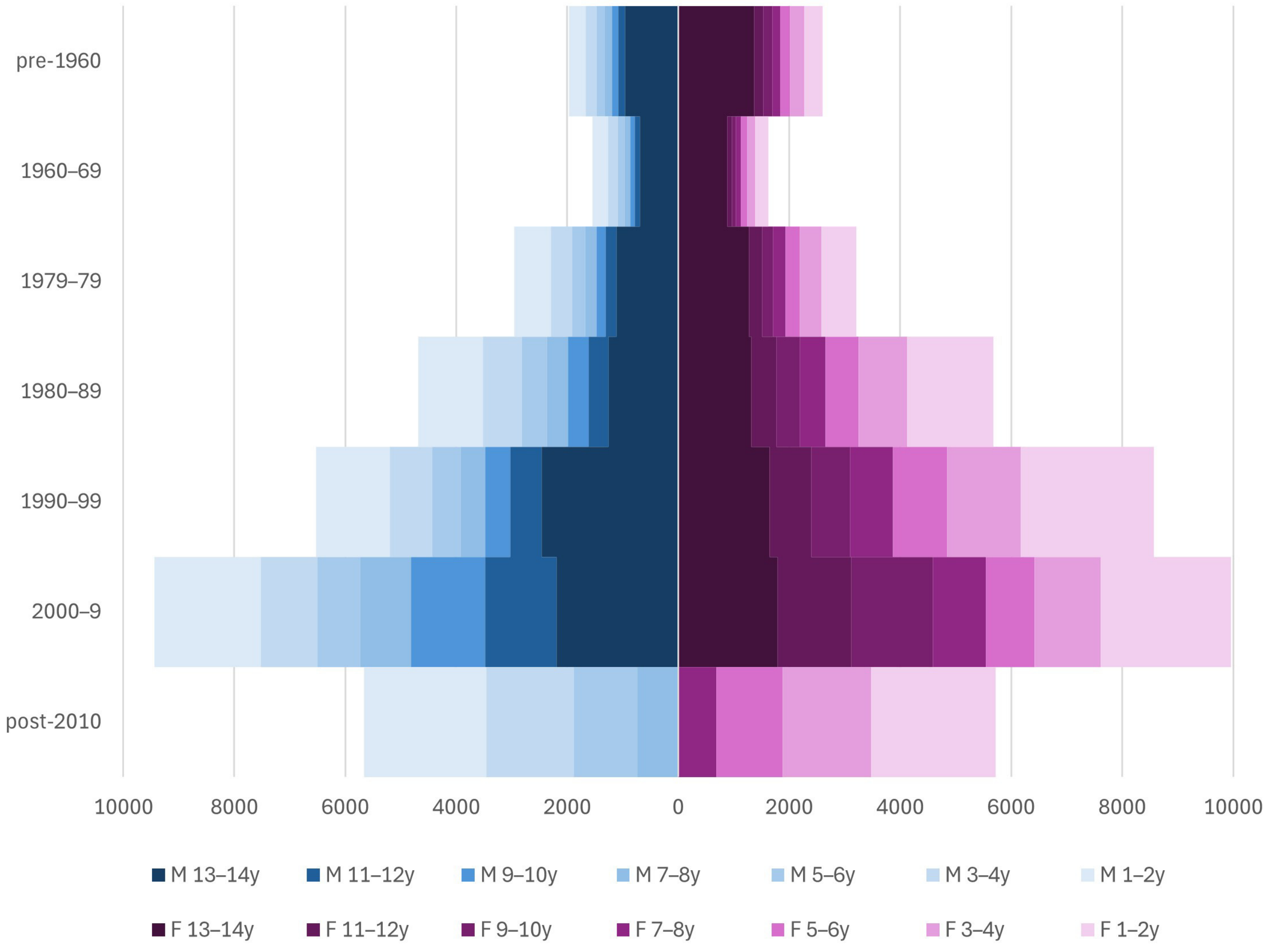

Figure 4. Number of HDSS residents by birth cohort and sex (males on the left, females on the right), and how many years they were present: darker shading indicates longer time in the HDSS.

Ethics

The Karonga HDSS has ethical approval from the outset of the Malawi National Health Science Review Committee (approval #20/11/2641, previously #416) and the London School of Hygiene and Tropical Medicine (approval #5081). All households provided written consent to participate in the Karonga HDSS, which could be rescinded at any time.

Data and software availability statement

Underlying data

Due to the detailed nature of the data describing the exact living arrangement of participants, it is not possible to anonymize it sufficiently in a way that allows it to still be useful; thus, the data are not available for open access. However, MEIRU welcomes requests to use data from bona fide researchers who should contact the first author (EM) in the first instance at info@meiru.mw. Full documentation of the dataset including a complete listing of variables can be found on the MEIRU data catalogue (https://kpsmw.lshtm.ac.uk/index.php/catalog/13) Further information on MEIRU datasets can be found on the MEIRU website.

Analysis code

Code is available through Zenodo: Family network and household composition: a longitudinal dataset derived from the Karonga HDSS, in rural Malawi (author-written code) https://zenodo.org/records/10037084

This project contains the following files:

- 0_master_KarongaHDSS_household_family.do: A Stata do-file which calls the following processing do-files.
- 1_identify_relatives.do: A Stata do-file in which all relative pairs are identified from parent and spouse id linkage lists for use in later do-files.
- 2_create_snapshots.do: A Stata do-file in which continuous HDSS residency episode data are reduced to snapshots.
- 3_assign_gps_to_snapshots.do: A Stata do-file in which GPS coordinates are added for each person for the household they are living in in each snapshot.
- 4_popdens_assign_hhmemb.do: A Stata do-file in which household summary variables are created, including population density and average relatedness within households.
- 5_id_household_members.do: A Stata do-file in which relationship between index and all household members are identified and summary variables created.
- 6_add_rels_10_250m: A Stata do-file in which index person's relatives within certain distances are identified and summary variables created.
- 7_get_other_datasets_ready.do: A Stata do-file in which other datasets related to socio-economic status and other factors are prepared for merging to the main dataset.
- 8_add_person_hh_states.do: A Stata do-file in which other datasets created in do-file 7 are merged to the main dataset and summary variables created.
- 9_combine_label_datasets.do: A Stata do-file in which all datasets are combined and labelled ready for use.

These files are available under the terms of the Creative Commons Attribution 4.0 International.

References

Bronte-Tinkew J, DeJong G: **Children's nutrition in Jamaica: do household structure and household economic resources matter?** *Soc Sci Med.* 2004; **58**(3): 499–514.
PubMed Abstract | **Publisher Full Text**

Crampin AC, Dube A, Mboma S, *et al.*: **Profile: the Karonga Health and Demographic Surveillance System.** *Int J Epidemiol.* 2012; **41**(3): 676–85.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Flinn MV, Quinlan RJ, Coe K, *et al.*: **Evolution of the human family: Cooperative males, long social childhoods, smart mothers, and extended kin networks.** In: Salmon, C.A., Shackelford, T.K. (Eds.), *Family Relationships: An Evolutionary Perspective.* Oxford Scholarship Online, 2007; 16–38.
Publisher Full Text

Gage AJ, Sommerfelt AE, Piani AL: **Household structure and childhood immunization in Niger and Nigeria.** *Demography.* 1997; **34**(2): 295–309.
PubMed Abstract | **Publisher Full Text**

Keilman N: **Recent trends in family and household composition in europe.** *Eur J Popul.* 1988; **3**(3–4): 297–325.
PubMed Abstract | **Publisher Full Text**

Koster J: **Family ties: the multilevel effects of households and kinship on the networks of individuals.** *R Soc Open Sci.* 2018; **5**(4): 172159.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Malawi Human Rights Commission: **Cultural Practices and their Impact on the Enjoyment of Human Rights, Particularly the Rights of Women and Children in Malawi.** 2006.
Reference Source

Maxfield MG: **Household composition, routine activity, and victimization: A comparative analysis.** *J Quant Criminol.* 1987; **3**: 301–320.
Publisher Full Text

McLean E, Dube A, Kalobekamo F, *et al.*: **Local and long-distance migration among young people in rural Malawi: importance of age, sex and family [version 1; peer review: 1 approved with reservations].** *Wellcome Open Res.* 2023; **8**: 211.
Publisher Full Text

McLean E, Price AJ, Palla L, *et al.*: **Data-driven versus traditional definitions of household membership and household composition in demographic studies: does latent class analysis produce meaningful groupings?** In: *International Population Conference.* Hyderabad, India, 2021a.

McLean E, Sironi M, Crampin AC, *et al.*: **Transitions to adulthood in rural Malawi in the 21st century using sequence analysis.** In: *International Population Conference.* Hyderabad, India, 2021b.
Reference Source

Perkins KL: **Changes in household composition and children's educational attainment.** *Demography.* 2019; **56**(2): 525–548.
PubMed Abstract | **Publisher Full Text**

Ponnighaus JM, Fine PE, Bliss L, *et al.*: **The Lepra Evaluation Project (LEP), an epidemiological study of leprosy in Northern Malaŵi. I. Methods.** *Lepr Rev.*

1987; **58**(4): 359–375.
PubMed Abstract | **Publisher Full Text**

Snyder AR, McLaughlin DK, Findeis J: **Household composition and poverty among female-headed households with children: differences by race and residence.** *Rural Sociol.* 2006; **71**(4): 597–624.
Publisher Full Text

Stiffman MN, Schnitzer PG, Adam P, *et al.*: **Household composition and risk of fatal child maltreatment.** *Pediatrics.* 2002; **109**(4): 615–621.
PubMed Abstract | **Publisher Full Text**

Strathman JG, Dueker KJ, Davis JS: **Effects of household structure and selected travel characteristics on trip chaining.** *Transportation (Amst).* 1994; **21**: 23–45.
Publisher Full Text

Tangeland T, Aas Ø: **Household composition and the importance of experience attributes of nature based tourism activity products - a Norwegian case study of outdoor recreationists.** *Tour Manag.* 2011; **32**(4): 822–832.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓ ✓

Version 2

Reviewer Report 21 June 2024

https://doi.org/10.21956/wellcomeopenres.23727.r81935

✓ **Sarah Reynolds**
University of California Berkeley, Berkeley, California, USA

Peer Review of "Family network and household composition: a longitudinal dataset derived from the Karonga Health and Demographic Surveillance System, in rural Malawi [version 2]"

Data Note

- Are sufficient details of methods and materials provided to allow replication by others?

Do files are available by request, so this code could be adapted if others had the raw data. I have not viewed the do files, so am unable to comment on the clarity of their annotations.

This is an exciting transformation of an administrative database to provide documentation of the relationships between all household members and their family proximity over time. Overall this work is thoroughly and clearly presented, and presents the possibility for an exciting expansion of the literature on family composition within and beyond the household longitudinally. I only have a few questions:

- Table 1 has data from every 4 years, Table 2 from every other year, and Figure 3 from every year. Improve consistency & explain in paper or footnote why these years were selected (except the one with every year)
- With what frequency is the household asset data collected?
- Understanding attrition is complicated. An illustration something like Figure 2 could illustrate ages at baseline and what percentage of participants are present at each year, with, of course 100% at baseline. This suggestion, however, leaves out individuals added at later years. Perhaps this is too complex to integrate and will be overwhelming – kudos if the author can think of a clever way to incorporate such information!

Highlighting some future directions can help researchers decide to use the data.

- For future applications, what variables link to external data sets? National ID number, cell-phone number, names?
- As a future extension, it could be noted that researchers could adjust snapshots to a particular relationship of interest. For example, if a researcher was interested in mother-child co-residence, the researcher could modify the do-files to focus only on this subset of

relationships and create continuous information rather than over 4 years.
    - Similarly, as an alternative to the binary approach for the presence of a relative at the snapshot point, an extension could be the percentage of time over the snapshot period during which the relative was present in the household.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Economics, Child Development, Family Demography, Intimate Partner Violence

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 03 June 2024

https://doi.org/10.21956/wellcomeopenres.23727.r83561

✓

Laurie DeRose
The Catholic University of America, Washington, District of Columbia, USA

The theoretical justification is clear: households matter in shaping and reflecting behavior. I get a clear picture of a complex data set. Even where I'm not thrilled with the process (identifying linked households), the immediate household are still preserved.
I like the way the files of the analysis code are organized.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Social demography

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 14 May 2024

https://doi.org/10.21956/wellcomeopenres.23727.r81007

✓

Ashira Menashe-Oren
Centre for Demographic Research (DEMO), Universite catholique de Louvain, Louvain-la-Neuve, Walloon Region, Belgium

Thank you for addressing my previous review. I have no further comments.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Demography

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 22 March 2024

https://doi.org/10.21956/wellcomeopenres.22594.r75859

? **Ashira Menashe-Oren**
Centre for Demographic Research (DEMO), Universite catholique de Louvain, Louvain-la-Neuve, Walloon Region, Belgium

This data note is a useful description of procedures in creating measures of household and kin structures with longitudinal data from a Health and Demographic Surveillance Site (HDSS). It is well-written and justified. I have two issues that I would like to see addressed (and a few minor comments for clarification or suggestions for improvement).

1. The data note informs on the process of generating variables describing household characteristics. It would be helpful if a list of each of these variables is noted up front at the very least. A table of summary statistics of some these variables (household size, genetic relatedness, presence of relatives within radii etc). Mostly, a few tables with examples of these variables seems essential – a couple of households with their members records, and then the various indicators (like one would see in Stata data browser). This would be especially helpful for non-Stata users since the code is only made available in Stata.
2. The section explaining the relationship between households is not clear, in particular the paragraph starting “Initially, a random sample of 100 pairs of households…”. I suggest revising this, and considering adding a flow-chart of the different steps taken.

Some minor comments:
Justification of using data from 2004 is needed, if the DSS was established in 2002.
The data is reduced to one record per person per period – taking a snapshot of mid-point. Does this not flatten/ simplify the data unnecessarily? One of the key advantages of the longitudinal data is the timing and order of events.
The parents or spouses who not reside in DSS but have identifying number – are they included in any of these household measures?
Is the proportion of unknown relationships truly an indicator of data quality? Is it not feasible for household members to not be related in Karonga?
Figure 2 could be for grouped years rather than single years.
Table 2 would be better as a population pyramid.
The data note ends rather abruptly. It would be nice to see a summary-like paragraph (before ethics section), relating back to the introduction.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Demography

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 05 Apr 2024

Estelle McLean

Thank you for your very helpful review. For your main 2 points: I have added in a table of some summary statistics, and I have added the link to the full documentation for the dataset which was inadvertently missing (table 1). I have tried to make the description of the methodology used for linking household clearer, and have added a diagram (figure 1). For your minor points: I have clarified in the text that the HDSS was set up from 2002-2004, so 2004 is the first complete year.

I have clarified the reason for reducing the longitudinal dataset to snapshots – the large number of variables and their rate of change (the combination of relatives living within 250 metres could potentially change very frequently) would mean that splitting the episodes and assigning these variables would be too computationally intense. Absent relatives are not included in this particular dataset, but could be identified if an analysis called for it. Unknown relationships refer to the missing identifiers – non-relatives may be identified if both have parent identifiers but no relationship link is found, I have clarified this in the text. I have not changed figure 2 to used grouped years, as using single years highlights the increase in unknown data in the last few years of the period. I have converted table 2 to a population pyramid figure (figure 4). I have added a summary paragraph before the ethics section.

Competing Interests: No competing interests were disclosed.