Demonstrating the value of Health and Demographic Surveillance Site data for complex secondary analyses, illustrated with analyses of young people's living arrangements and transitions to adulthood.

ESTELLE MARIE MCLEAN

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy of the University of London

JANUARY 2024

Department of Population Health
Faculty of Epidemiology and Population Health

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Supervisors:
Rebecca Sear and Emma Slaymaker

Research group affiliation:
Malawi Epidemiology and Intervention Research Unit

Funding:
No funding was received

I, Estelle McLean, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Estelle McLean
22 January 2024

Abstract

Background

Health and demographic Surveillance Sites (HDSS) are long-running geographic cohort studies where detailed data are regularly collected on residents. HDSSs are valuable resources of longitudinal data but their complexity can present barriers for fully utilising them.

Objectives

This thesis aims to demonstrate the value of using complex data manipulation and analytic techniques on secondary data from the Karonga HDSS, Malawi. Two key attributes of this data resource are exploited, repeated data points and detailed household and family links, to present analyses related to adolescents: who they live with and their transitions to adulthood.

Methods

Complex longitudinal datasets were generated, with variables indicating presence of family. Two data reduction techniques were utilised: latent class analysis, to generate data-driven household composition variables, to compare against 'standard' definitions; and sequence analysis, to explore trajectories in transitions to adulthood. These analyses inputted into two longitudinal analyses: an analysis of how presence of family is associated with migration in young people; and an analysis of divorce at a young age and its impact on transition to adulthood.

Results

The data reduction techniques added value to HDSS data analyses to improve understanding of complex phenomena: latent class analysis demonstrated the breadth of household types, and sequence analysis showed the diversity in adulthood trajectories, and both techniques enabled useful classification into clusters. They were also useful as exploratory tools, for generating variables and ideas. The resulting analyses enabled detailed understanding of the lives of young Malawians, including key differences by sex, i.e. female adolescents were more likely to migrate than males and at an younger age; and women were more likely than men to have their transition to adulthood disrupted following a divorce if they had children.

Conclusions

HDSSs represent rich resources with great potential for complex secondary analyses and greater flexibility for data manipulation compared to other longitudinal surveys. This complexity also means that great care is needed to ensure that appropriate manipulations and techniques are used.

List of tables

List of figures

List of commonly used abbreviations

aOR	Adjusted odds ratio
aHR	Adjusted hazard ratio
BIC	Bayesian information criterion
CA	Cluster analysis
CI	Confidence interval
DHS	Demography and Health Surveys
HDSS	Health and demographic Surveillance Site/System
HH	Household
LCA	Latent Class Analysis
MEIRU	Malawi Epidemiology and Intervention Unit
SES	Socio-economic status

Table of Contents

1. Introduction

This thesis is an exploration of how longitudinal data from a Health and Demographic Surveillance site (HDSS) can be used to investigate questions of relevance to health and demography which may not have been planned when the systems were designed. In this introduction, I first provide background on HDSSs, and the specific site which I use in all my analyses: the Karonga HDSS in northern Malawi. I use the widely used Demographic and Health Surveys as a comparison to highlight important strengths and weaknesses of HDSS data, including issues around sharing data. Finally, I introduce the theme which I have used in this thesis to demonstrate uses of HDSS data for secondary analyses: adolescence and the transition to adulthood.

1.1. Health and Demographic Surveillance Sites

Health and Demographic Surveillance sites (HDSS) are data collection exercises which operate in geographical areas, which may be one contiguous area, or multiple separate ones. Everyone living within the defined area is eligible to take part. Participants may enter the HDSS system through the full census carried out at the start of operations, birth or moving into the area, and may leave the system through death, moving out of the area, or if the HDSS stops operations. HDSSs record all participants in household groupings; the definition of a household may vary slightly from area to area, but often requires recognising the same head and ‘eating from the same pot’. Some HDSSs allow for absent members, i.e. those working away from home but still considered part of the household. The data can be used to monitor trends in mortality, fertility and migration: HDSSs exist in low and middle income (LMIC) countries where they have been described as a short to medium-term solution to filling the gap of vital registration systems (Ye et al., 2012).

As well as collecting data on demographic events, HDSSs tend to regularly gather other health and socio-demographic data from the participants, and they are also used as platforms for specific studies which make use of the existing infrastructure to firstly select and find the specific type of participant they need, and then to find them again, if the study requires multiple interactions. As HDSSs only operate in specific small geographic areas, there have been concerns over the generalisability of the data and results produced, however it has been argued that this can be mitigated to some extent through triangulation with other sources, and that the richness and quality of the data and statistical power are more important when considering their usefulness (Bocquier et al., 2017b). They tend to

capture data for many years: the oldest HDSSs were set up in the 1940s and 50s (Herbst et al., 2021; Ye et al., 2012); meaning that they are rich sources of data for secondary longitudinal analyses, for example this review of analyses made possible by several decades of demographic data collection in Kiang West, Gambia (Moore, 2020).

1.2. Issues with sharing HDSS data

Production of HDSS data takes a huge amount of resources from study design, programming, logistics, field-work, data entry, cleaning and storage and can represent quite a burden for the study population who give a lot of time to provide the data. The resulting datasets represent incredibly rich resources which should be used to their full potential. Along with other public health datasets there have been calls to make HDSS data more accessible to external scientists (Chandramohan et al., 2008). Researchers often struggle to share their data as they need return on their investment of the time spent collecting the data, and the additional time, resource and skill/expertise needed to generate suitably documented datasets do not tend to be recognised in academic career progression pathways (Chawinga and Zinn, 2019): there have been calls to reduce some of these barriers by increasing recognition of the vital role of data professionals (Pisani and AbouZahr, 2010). In addition to standard barriers to data sharing, HDSS sites are exclusively operated in low income settings where both analytical and data management capacity are lower and there is an understandable reluctance to professionals from low income settings being expected to collect and curate datasets, with little recognition and career advancement, only for the data to be analysed by researchers from high income settings (Chandramohan et al., 2008; Hinga et al., 2021). Additionally, the complexity of the HDSS datasets requires quite a lot of understanding to analyse to avoid accidentally introducing biases or drawing the wrong conclusions, and with multiple records per participant it is also difficult to share data in an open access form, as suitably anonymised data may become useless for analysis.

Despite all the issues detailed above, some HDSS data is available openly. The main platform for HDSS data is the INDEPTH network (https://www.indepth-ishare.org/index.php/home). This is a collaboration of multiple HDSSs across the globe providing analytical support and workshops, as well as a platform to share basic harmonised HDSS datasets (Herbst et al., 2015a). These datasets are available to external researchers, and while they do not contain all the possible data from each HDSS, have been used for multiple pooled analyses. Additionally the Mekong HDSS which ran from 2001 to 2006,

released their full datasets open access in 2016 (https://www.icpsr.umich.edu/web/DSDR/studies/36601/publications) (Heuveline et al. 2016), however there do not appear to be any publications which used it. After 2006 the project was continued and expanded, but more current data does not appear to have been released (Heuveline et al. 2017). The Agincourt HDSS in South Africa also makes some of their data openly available: they regularly update their '1 in 10' dataset, which is a 10% sample of the full dataset (https://www.agincourt.co.za/data). Anyone is free to download this sample data, to plan analyses and then a full request with individual approval is needed to obtain the full set. The Karonga HDSS contributes data to the INDEPTH network and has shared data with many individual researchers and collaborations.

Given the issues with directly sharing HDSS data, I believe the way to increase accessibility of HDSS data is through fair collaborations, rather than attempting to share open access anonymised data. I also believe that there is much scope for complex and in-depth analyses using HDSS data and was keen to increase visibility firstly by demonstrating the complex and interesting ways that the data can be used but also by exploring the potential limitations and pit falls of using some data techniques on these datasets.

1.3. Demography and Health Surveys vs. Health and Demographic Surveillance Sites

Demography and Health Surveys (DHS) are a useful comparison for HDSS data. DHS is a program of nationally representative surveys carried out in low and middle income countries. It started in 1984, developing from World Fertility Surveys and contraceptive prevalence surveys in the 1970s and 1980s (Cleland, 2010). The core surveys collect information on a range of population and health topics, and optional modules are available for countries to add if required (Fabic et al., 2012). DHSs are meant to be carried out regularly, usually every 3-6 years, though individual countries decide upon the timing and frequency of the surveys. Sampling is stratified by geographic area and, depending on the size and needs of the country, 5000-30,000 households are included. At each household, a 'roster' of household members is recorded, along with a household survey and then surveys carried out with specific individuals, usually all women aged 15-49 and their children aged 0-59 months, and often men aged 15-59 years (Corsi et al., 2012).

One of the great strengths of the DHS data is that it is all freely available via their website, either as individual data, or summary statistics and graphs using web-based analytical tools

(https://dhsprogram.com/). This means that published analyses of DHS data are very common, and have been increasing yearly since the inception of the surveys, most commonly analyses of fertility, family planning, sexual behaviour or maternal and child health or nutrition (Fabic et al., 2012). A review of uses of DHS data in the literature produced a varied list of types of analyses, including examining trends in health outcomes, comparative analyses of associations between exposures and diseases in multiple countries, geographical variation in nutritional status, associations between maternal or paternal exposures and child outcome, comparisons in child outcomes (i.e. nutritional status) before and after economic events or introduction of policies (Corsi et al., 2012). Similar analyses would be possible with HDSS data so, given the availability of this powerful resource, a question might be asked over the comparative value of HDSS data.

In table 1.1 I have summarised the advantages and disadvantages of both platforms, much from the DHS side is from a 2012 profile of DHS data (Corsi et al., 2012). The DHS has advantages over HDSS in terms of national representativeness: only very small geographic areas can be included in HDSSs, however the full population nature of HDSSs (rather than samples) may be an advantage for some kinds of study: i.e. data on full clinic catchment areas are available. Repeated DHS surveys allowed for assessment of population-level trends over time, however there is no information in years between DHS surveys, and, as new participants are interviewed for each survey, no possibility of conducting longitudinal analyses of individuals, or following them up for further surveys. From this perspective, the HDSSs have advantages, as continuous data are available for all years and individuals are linked across all time points: individuals longitudinal analyses are therefore possible, and HDSSs can be used as sampling frames for further studies (the value of HDSSs as sampling frames is well described in an article suggesting their value in microbiome research (Agarwal et al., 2017)).

The HDSSs are at an advantage in terms of linking between individuals as well: linkages are usually possible between household members, parents, children and spouses, and in many HDSSs it may be possible to link HDSS members to other data sources, such as clinic records. Within the DHS, mothers may be linked with their children, and their spouses, if they are included. However, often only women of child-bearing age are included in DHS. This latter point is probably the most important with regards to interpreting DHS data, as infant and child mortality rates can only be calculated based on data from live mothers in that age group, so there is a potential for bias if maternal survival is linked with infant mortality.

A key advantage of DHS data over HDSS is the consistency and comparability of methods and data across countries. DHS was set up with harmonisation and pooled analyses in mind, the continuous funding and support allows researchers across the world to access and analyse the data. HDSSs do not have such an over-arching structure of support and funding, and while there have been attempts to pool and harmonise the data, the core datasets generated by each site are often quite different. The longitudinal nature of the HDSS data make it more complex for analyses: a greater level of statistical and data manipulation expertise is required for HDSS analyses compared to DHS ones.

HDSSs have been used on several occasions to validate DHS data, either by conducting a DHS-style survey within the HDSS area and comparing the results from HDSS data from the same households (Bairagi et al., 1997; Jasseh et al., 2022) or by comparing results from the actual DHS survey with HDSS data (usually just using DHS results from the specific are or district the HDSS is in) (Deribew et al., 2016; Fottrell et al., 2010), these studies tended to find that fertility rates were similar from the two sources, but that mortality rates may be over-estimated in the DHS data compared to the HDSS. These comparisons show that both DHS and HDSS complement each other for producing useful analyses on demography and health.

Table 1.1: Comparison of advantages and disadvantages of DHS and HDSS

Demography and Health Surveys (DHS)		Health and Demographic Surveillance Sites (HDSS)	
+	Nationally representative	-/+	Only small area and small portion of population included, however the entire population is included
+	Data available on-line	+/-	Some data is available, i.e. through the INDEPTH network, most HDSS data is only available through requests to specific sites
+	Harmonised data available across countries. Common data standard exists	+/-	Each site's data is different and requires additional effort to harmonise for pooled/comparative analyses. Only a limited common data standard/model
+	Comparative data capture techniques across countries	+/-	Procedures and definitions may vary across sites, however they are relatively similar
+	High response rates	+	Very high response rates
+	Repeated cross-sectional surveys allow analysis of change at population level	+	Longitudinal data allows analysis of change over time at population level
+	Mother data can be linked to child data and cohabiting couples can be linked if both name the other.	++	Data on individuals can be linked to other household members, usually parents and often other relatives
-	Surveys cannot be linked to analyse change over time at individual level	+	Longitudinal data allows for life-course analysis of individuals
-	Surveys in different countries happen at different times with different frequencies	+	Data are continuous so any year can be looked at
-	Only certain individuals included, i.e. women aged 15-49	+	All ages included
-	It is not possible to identify and follow-up individuals	+	Can act as a sampling frame for follow-up studies
-	Lots of data are proxy reported	-	Proxies also often used
+	Little/no data cleaning required	-	Data cleaning required, especially to deal with inconsistencies with longitudinal data, and expertise with the HDSS required for making appropriate data cleaning decisions.

1.4. The Karonga HDSS and inspiration for this PhD

The Karonga HDSS was set up in 2002-2004 and has been running ever since, albeit with a period of 'skeletal' running from 2018-2021 due first to funding issues, then the Covid-19 pandemic, when only births and deaths were recorded through phone calls, as face-to-face data capture was suspended. The population was initially around 30,000 and in 2022 was over 47,000. The area covered is 135 km^2 in the south of Karonga district in the north of Malawi. While HDSSs are similar in the type of data collected, the way these data are gathered vary. The majority of HDSSs visit every household at least annually (some up to 3 times a year) to collect data on births, deaths and movements in and out of the household since the last visit. Karonga HDSS is slightly different, in that births and deaths are reported by 'Key informants' on a monthly basis. These are community members who are given a small stipend and asked to monitor up to about 40 households. Once they report a birth or death, trained fieldworkers visit the household in question to fill in the required forms (Crampin et al., 2012).

The Karonga HDSS was set up in an area where demographic and health research had been going on since 1979: work initially focussed on understanding and control of leprosy, which involved district wide full household surveys. As leprosy runs in families there was a lot of work dedicated to linking participants to their mother and father identifiers, so family trees could be constructed. This linkage work continues for HDSS members, and additionally linkages between spouses are also made. Every household has GPS coordinates recorded when registered.

I have been working with the Karonga HDSS since 2014. In my role as the data scientist, I am involved in all aspects of data from designing of studies, implementation of data capture, storage and documentation of data, creation of analytical datasets and the analyses themselves. Working closely with the data, I could see that the possibilities for secondary analyses were broad. I chose to take the opportunity to do a PhD to demonstrate some of these.

1.5. Children, adolescents and the transition to adulthood: existing research in Malawi

The choice to use adolescents and the transition to adulthood as the subject for the analyses for this PhD was made based on 4 factors. Firstly, the family linkage data with the Karonga

HDSS datasets are better for younger people so there would be less need to drop records due to missing data; secondly, adolescence is a key period of transitions and changes so is ideal to demonstrate longitudinal data methods; thirdly, analyses of adolescence and the transition to adulthood have public health importance, as experiences during these times can have effects on health and well-being both in the short-term and later into life; and finally there was already a sizable body of research on adolescents in Malawi to draw on, to compare and contrast findings.

A detailed national survey of adolescents aged 12-19 was carried out in 2004; this covered areas such as living arrangements, schooling, work, puberty and initiation ceremonies/rites, relationships, sexual health, pregnancy and child-bearing, and risk behaviours (Munthali et al., 2008). Of note for my analyses, they found that found that the national median age at menarche was 15.1 and the median age for first sign of puberty for boys was 14.6; they found that girls in the Northern region (where the Karonga HDSS is located) tended to start menstruation earlier than girls from the Southern and Central regions. This survey also reported in detail on initiation ceremonies and rituals experienced by adolescents: they reported that in the Northern region these are virtually non-existent for boys, and for girls tend to be educational sessions through the church or village elders (Munthali and Zulu, 2007).

There have been 3 key longitudinal studies involving adolescents in Malawi:

1. ‘Malawi Schooling and Adolescent Study’ which followed up a sample adolescents from 2 districts in Southern Malawi who were aged 14-16 in 2007 through 5 rounds of data collection up to 2013 (Hewett and Mensch, 2019). Outputs from this study include one studying migration where they found high levels of mobility amongst female and male adolescents both for marriage and economic opportunities (Chalasani et al., 2013), an analysis which found that parental death, but not divorce, was associated with young women’s own divorce (Grant and Pike, 2019), one examining the effect of perceived future risk of HIV on marriage and divorce (Grant and Soler-Hampejsek, 2014) and the association of school attainment on later paid or unpaid work for women and men (Soler-Hampejsek et al., 2021).

2. The Marriage Transitions in Malawi project in the central region followed up a cohort of adolescents up to 5 times over a 2 year period (Beegle and Poulin, 2017), and have analysed migration as part of the transition to adulthood (Beegle and Poulin, 2013), pre-

marital fertility as a predictor of marital age (Poulin et al., 2021) and household income shocks as a predictor of marriage (Molotsky, 2019).

3. The Tsogolo la Thanzi (Healthy Futures) study was a longitudinal follow up (2009-2014) of young Malawians aged 15 to 25 at baseline, the focus was on relationships in the context of HIV, collecting data on fertility intentions, relationships, sexual behaviour and health (Yeatman et al., 2019). Outputs have included analyses assessing changing fertility desires (Sennott and Yeatman, 2012) and trajectories in family planning within a relationship (Furnas, 2016).

Other research including adolescents includes a study in Northern Malawi found that orphanhood was a strongly predictor of age at sexual debut (Mkandawire et al., 2013) and several studies focussing on adolescents' and HIV (Kim et al., 2017; Mandiwa et al., 2021). There have also been high quality examples of qualitative work on adolescents in Malawi, including from perspective of the adolescents (Zietz et al., 2018) and of their parents (Grant, 2012), and specifically about the transition to adulthood (Kok et al., 2021).

This body of work gave a good background for me to triangulate my findings against, but also showed that there were still plenty of places to add to the literature: the longitudinal surveys took place in the southern and central regions while the Karonga HDSS is in the North, and many of the analyses focus on girls/young women while only a few include boys/young men. The Karonga HDSS data has already been used to look in-depth at the associations between age at menarche, sexual debut, early pregnancy and school performance and drop-out (Glynn et al., 2018, 2010; Sunny, 2018; Sunny et al., 2019, 2018, 2017). I felt there was still areas to consider with our data: including looking at the transition to adulthood more holistically by including multiple outcomes simultaneously, including outcomes that had not previously been examined (i.e. leaving home) and including boys/men. I also planned to make greater use of the key features of the HDSS: the longitudinal data and links within families.

2. Research objectives

My overall objective for this thesis is:

To demonstrate the use of complex data manipulation and reduction techniques on existing HDSS data to usefully answer questions related to health and demography.

This can be broken down into several objectives:

1. Data objectives
    - 1a. To demonstrate the utility and value in complex secondary analyses using HDSS data not originally collected for these purposes
    - 1b. To assess the value and potential disadvantages of certain data manipulation and data reduction techniques for use with HDSS data
    - 1c. To increase knowledge and visibility of HDSS datasets and the scope of potential secondary analyses

2. Adolescent objectives
    - 2a. To describe who rural Malawian adolescents are living with, and how often, and how, these living arrangements change; and to assess whether these changes affected by presence of family outside the household.
    - 2b. To describe how rural Malawians experience transitions to adulthood and assess whether this has this changed over time
    - 2c. To investigate how divorce at a young age affects the transition to adulthood for young rural Malawians.

3. Thesis summary

Below I briefly introduce the remaining chapters in the thesis.

3.1. Chapter 4: Use of Health and Demographic Surveillance Site data for secondary analyses: guidance for researchers following a review of existing analyses

HDSSs represent rich sources of data for analyses, and their structure allow for quite complex data manipulations to answer research questions they were not originally designed for. As they are open cohorts, they also have issues which can make interpretation of data and conclusions complex: namely that lack of data on in- and out-migrants when they are outside of the area may cause bias for certain outcomes and that data repeatedly collected may result in inconsistencies and/or more reliable data for older participants or those present in the HDSS for more time. HDSS analyses are used to further research and policy agendas so analyses must be appropriately conducted, and also reported in such a way that these issues, and the effect they may have had on the conclusions, should be clearly stated allowing clear understanding by people not involved in HDSSs.

This chapter presents the results of a systematic review of the literature which I carried out to identify examples from other HDSS which leveraged HDSS linkages between individual time points, or between separately individuals to assess how other researchers had dealt with data issues. I reviewed the analyses using 6 key aspects of conduct and reporting of HDSS analyses and present the findings with recommendations to new and existing users of HDSS data. This chapter will be submitted to a peer reviewed journal.

3.2. Chapter 5: Family network and household composition: a longitudinal dataset derived from the Karonga HDSS, in rural Malawi

I spent quite a long time designing my data manipulation methods to generate my core datasets which I then adapted for my individual analyses. I had to make some decisions to balance the computer memory needed to perform intensive data manipulation on multiple records (i.e. calculating physical distance between each index person and multiple relative types) with having the most complete and detailed dataset as possible. I was used to using the HDSS data as a continuous dataset, splitting each person's follow-up time into new periods if an exposure changed, however as I had multiple exposures (living with each relative type, living with 50m of each relative type etc.) I could have ended up having to split

person-time into episodes of very short duration resulting in datasets that were too large to manage with the tools I had access to. I decided to reduce the data into snapshots (mid-year (over 500,000 records) or mid-quarter (over 2 million records) depending on the analysis) in the hope that the loss of data would be made up for in the richness of the exposure variables that I could create. In this methodology chapter, I describe the processes used to create my base dataset and the variables created. The chapter is in the form of a 'data note' a description of a dataset published either as a stand-alone document, or to complement an analytical paper. I will submit this data note to Wellcome Open Research.

3.3. Chapter 6: Data-driven versus traditional definitions of household membership and household composition: does latent class analysis produce meaningful groupings?

Across all HDSSs, members are grouped into households. This aids field and data management as people are assigned to the place where they are most likely to be found, and means that some data for a group only need to be collected once rather than for each individual. From an analytical perspective, household groupings are used to allow for clustering and to generate summary variables. To avoid double-counting in most HDSSs, and other household surveys, participants may only be assigned to one household, however this may not actually reflect the way people live. People may actually contribute to, or feel part of more than one, i.e. children of divorced parents living across 2 households, or an adult working away but sending money back home. In the Karonga HDSS, men who are married to more than one woman (which is relatively common in the area) are assigned to each wife's household, however other participants are only assigned to one. Some other HDSSs record whether households are living together in a compound, this is not available in the Karonga HDSS data, however many families in the area do live close together and are likely to share some resources. Household composition (the relationships between the members) is often used in sociological and demographic analyses, however the 'standard' categories such as nuclear and extended family may not be appropriate for an African setting.

For this analysis I used household GPS data to link households into expanded groups (as a proxy for compounds), and then used a data driven technique to develop household composition variables: Latent Class Analysis (LCA). I used the latter as I was aware of my position of a foreigner in Malawi and was keen to avoid bringing my own biases into the process. In the chapter I display in detail how I explored the data with LCA, and how I developed household composition variables based on the LCA results. I then explore the

household composition variables and the effect of using an expanded household membership definition on some example analyses. This chapter is in the format of a research article, which is currently under review at the Journal of Biosocial Science.

3.4. Chapter 7: Investigating use of sequence analysis to assess changes in transition to adulthood over time using HDSS data

Adolescence is a key period of development and experiences during this time can have an effect well into adulthood. Becoming an ‘adult’ is both a biological and a social process, the latter usually involving some or all of several milestones and markers, such as leaving school, becoming financially independent, marrying and having children etc. These markers and transitions may often be studied separately. However, firstly, they may linked (i.e. leaving school to get married) and secondly, transitions may not be smooth or one-way, with young people leaving and returning home, entering and ending relationships etc. These factors encourage the use of methods which use multiple linked records for individuals over time, such as sequence analysis. This method treats individuals’ experiences as text strings, where each letter represents a ‘state’ for a certain period of time, for example “NNNUUUcUcUcDDD” could represent a person spending 3 periods as ‘never married’, 2 in a ‘union’, 3 in ‘union with child’ and 3 as ‘divorced’. The process assesses how similar or different sets of sequences are, and then cluster analyses can be used to assign people into groups. This technique has been used previously to understand the transition to adulthood in both high and low income settings, however it has only rarely been used on HDSS data (for any outcome).

In this chapter I explore using sequence analysis to study the transition to adulthood using HDSS data. A key issue with HDSS data which make this kind of analysis difficult is that information is not available on people when they are outside of the HDSS area. This is particularly an issue for studying the transition to adulthood, as, for adolescents, migration is likely to be related to adulthood markers as they may move for school, for marriage or for divorce. I describe 3 approaches at sequence analysis including assessments of the effects of missing data due to migration, and reflect on the utility of the method for use with HDSS data.

3.5. Chapter 8: Local and long-distance migration among young people in rural Malawi: importance of age, sex and family

Migration is common among young people in sub-Saharan Africa and for older adolescents and young adults. Studying migration is important on a population level to understand population changes and needs, and also on an individual level as it can be associated with both positive and negative outcomes: i.e. migrants may have greater access to education and work opportunities, but in some studies migration has been found to be associated with risk behaviours. For young people, local and long distance moves may be related to markers of adulthood, such as schooling, marrying or getting divorced, so studying it helps to give insight into the transition process. HDSS data is also well suited to studying migration: in- and out-migration are key HDSS events regularly collected as part of the core processes, so the data are likely to be reliable.

In this analysis I examine migration in both children and adolescents/young adults. This was the only analysis where I included younger children, this was because their movements provided useful information about the lives of their parents. The linkages between household members and relatives within the HDSS data allowed me to distinguish between ‘accompanied’ and ‘independent’ moves, as the latter would provide more information about the transition to adulthood. I use the household composition variables derived in the LCA chapter and present both descriptive and regression analyses, the latter using multi-level modelling. This chapter appears in the format of a research article which is currently under review on the Wellcome Open Research platform.

3.6. Chapter 9: Divorce and the transition to adulthood in rural Malawi

Both the sequence analysis and the migration analysis suggested that divorce played a role for some of the young Malawians analysed. Divorce is common in Malawi, often occurring within the first 3 years of a marriage. Anecdotally in Karonga, the end of a union between 2 people who married very young may be seen as a positive step, as it allows them to return to school and improve their prospects. However, data from other settings has indicated that marriage may be chosen because continued schooling is not an option, rather than marriage precipitating school drop-out, implying that divorce would not necessarily lead to higher education. Additionally, it has been suggested that young men may be relatively unaffected by a divorce, however for women, having children is more likely to leave them disadvantaged with regards to both remarrying and going back to school.

In this analysis I investigate the effects of gender and child-bearing on divorce and remarriage at a young age, and also the impact of divorce on other markers of adulthood (living with family and attending school). This analysis uses the transitions dataset used for the sequence analysis to identify first marriages, marital disruption and remarriages; however fewer people are lost from the analyses as fewer continuous data points are required for inclusion. Standard survival and logistic regression analyses are used, the former allowing people who migrate to contribute time to the analysis when they are present, and the latter focussing on single cross-sections of the data, only including people who were present at that time point or age. This chapter appears as a research article which I intend to submit to a peer-reviewed journal.

3.7. Chapter 10: Discussion and conclusions

Following a brief summary of the findings of the chapters 4-9, which are described in relation to the thesis objectives, I discuss my findings related to adolescence and the transition to adulthood in the context of existing literature, and identify areas for future work. I then reflect on the experiences I had with the Karonga HDSS data, describe current areas of development and identify gaps, drawing upon the same framework I used to review HDSS paper in chapter 4: data manipulation, dataset structures, statistical methods, repeated measures, migrations and missing data. Many aspects discussed in this section apply to other HDSSs and the wider HDSS community. Finally, I discuss how my work might be possible using data from other HDSSs.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	145184	**Title**	Ms
First Name(s)	Estelle		
Surname/Family Name	McLean		
Thesis Title	Demonstrating the value of Health and Demographic Surveillance Site data for complex secondary analyses, illustrated with analyses of young people's living arrangements and transitions to adulthood.		
Primary Supervisor	Rebecca Sear		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Not yet decided
Please list the paper's authors in the intended authorship order:	Estelle McLean, Emma Slaymaker, Rebecca Sear
Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I developed the concept, designed and carried out the literature review, reviewed and summarised the papers, wrote the first draft and redrafted following comments from co-authors

SECTION E

Student Signature	
Date	27 July 2023

Supervisor Signature	
Date	27 July 2023

4. Use of Health and Demographic Surveillance Site data for secondary analyses: guidance for researchers following a review of existing analyses

Estelle McLean[1,2#], Emma Slaymaker[2], Rebecca Sear[2]

1. *Malawi Epidemiology and Intervention Research Unit, Lilongwe, Malawi.*
2. *London School of Hygiene and Tropical Medicine, Faculty of epidemiology and population health, London, United Kingdom*

 # author for correspondence: estelle.mclean@lshtm.ac.uk, ORCID: 0000-0001-6079-0663, London School of Hygiene and Tropical Medicine, department of epidemiology and population health, Keppel St, London WC1E 7HT, United Kingdom

4.1. Abstract

Health and Demographic Surveillance sites (HDSS) are geographic open cohorts designed to monitor trends in mortality, fertility and migration in countries with absent or incomplete vital registration systems. Data on demographic events, socio-demographic indicators, and often, certain health conditions are gathered on the whole population, often for decades. Participants are grouped into households, and often links between parents and spouses are present, allowing for complex longitudinal analyses which may take household and familial contexts into account. As they are open cohorts they also have issues which can make interpretation of data and conclusions complex: namely that lack of data on in- and out-migrants which they are outside of the area may cause bias for certain outcomes and that data repeatedly collected may result in inconsistencies and/or more reliable data for older or participants present in the HDSS for more time. This article describes and presents results of a systematic review of the literature carried out to identify examples from other HDSS which leveraged HDSS linkages. Selected papers were assessed for how the researchers conducted and reported their analysis according to 6 important aspects as HDSS analysis. Findings are presented to describe good and less optimal practices, along with recommendations to new and existing users of HDSS data.

4.2. Introduction

Health and Demographic Surveillance sites (HDSS) are geographic open cohorts: everyone living within a defined area is eligible to take part. Participants may enter the HDSS system

through the full census carried out at the start of operations, birth or moving into the area, and may leave the system through death, moving out of the area, or if the HDSS stops operations. Readers are directed to other recent articles which give a useful and detailed descriptions of the past present and future of HDSSs (Herbst et al., 2021) and important ethical issues (Ghafur et al., 2020; Hinga et al., 2021). HDSSs are designed to monitor trends in mortality, fertility and migration: they have been described as a short to medium-term solution to filling the gap of vital registration systems (Ye et al., 2012). As well as collecting data on demographic events, HDSSs tend to regularly gather other health and socio-demographic data from the participants, meaning that they can also be used for other longitudinal analyses. Individual and pooled analyses of HDSS data on various subjects can be found in the literature, for example mortality (Wasko et al., 2022); HIV (Roberts et al., 2022); religiosity (Lynch et al., 2022); education and poverty (Werner et al., 2022); nutrition (Mank et al., 2021) and vaccination (King et al., 2020). HDSSs are relatively responsive systems and the existing infrastructure can be used to respond to public health issues, for example existing and new HDSS processes (i.e. linkage of HDSS records with clinic records) and data (i.e. new capture of behavioural and biomarkers data) were leveraged to inform HIV policy (Herbst et al., 2015b) and several HDSS used existing and new systems to respond to the Covid-19 pandemic (Romero Prieto et al., 2022; Siedner et al., 2021).

Some HDSS data is available via the INDEPTH network (https://www.indepth-ishare.org/index.php/home). This is a collaboration of multiple HDSSs across the globe, providing analytical support and workshops, as well as a platform to share basic harmonised HDSS datasets (Herbst et al., 2015a). These datasets are available to external researchers and, while they do not contain all the possible data from each HDSS, have been used for multiple pooled analyses. Two training manuals are available; firstly, describing how to obtain data from the INDEPTH system, how to format it for event history analysis/survival analyses and how to carry these out (Bocquier et al., 2017a) and secondly, specifically on migration and mortality analyses (Bocquier et al., 2019). Other data can be accessed through collaborations with individual HDSS sites, or through other networks such as the ALPHA network (Reniers et al., 2016). Individual HDSSs tend not to share their data open access. Firstly, anonymising the full complex dataset sufficiently to preserve the privacy of participants yet still allowing useful analysis is complex (Templ et al., 2022). Secondly, HDSS sites are exclusively operated in low income settings where both analytical and data management capacity are lower and where there is an understandable reluctance to professionals from low income settings being expected to collect and curate datasets, with little recognition and career advancement, only for the data to be analysed by researchers from high income settings (Chandramohan et al. 2008; Hinga et al. 2021). Equally, however,

due to the high cost of gathering the data in monetary terms, as well as in terms of the burden on participants and HDSS staff, it is important that the data are used to their full potential.

HDSSs tend to capture data for many years: the oldest HDSSs were set up in the 1940s and 50s (Herbst et al., 2021; Ye et al., 2012); meaning that they are rich sources of data for secondary longitudinal analyses. They also collect data continuously, meaning that data are available for all years that the HDSS is in operation. Participants of all ages are included in HDSSs and they are grouped according to their household, which often have GPS coordinates recorded. Often linkages between parents and children, and between spouses are available. This means that powerful analyses taking environmental, household and family contexts into account can be carried out. As HDSSs only exist in small areas, they are not generalisable to the wider population, however they include the full population in a specific area and tend to involve large numbers of participants, which are advantages for certain analyses (Bocquier et al., 2017b). HDSSs exist in many low and middle income countries across the globe, and while processes and dataset are not fully harmonised, they can be used for pooled or comparative analyses. Considering all HDSSs, coverage of different types of ecological and environmental areas, and socio-economic strata has been found to be relatively high (Edson Utazi et al., 2018; Jia et al., 2015; Tatem et al., 2006). The populations under observation may become different because of the HDSS, and participation fatigue may affect data (Herbst et al., 2021). As HDSSs are open cohorts, some participants may have fewer data points than others if they migrate in or out of the area, and migration may be related to outcomes (Bocquier, 2016). Equally some participants will have data collected repeatedly which may lead to more complete and accurate data but may also lead to inconsistencies. These may become more important as more complex data linkages and manipulations are carried out, as people with fewer data points may need to be dropped which can lead to biases, and people with more data may have more reliable data than newcomers, which may also affect analyses.

Like most health and population research, HDSS data is generated with one of the goals being to influence local, national and global policy (Herbst et al., 2015b; Williams et al., 2010). HDSS analyses are also commonly used in systematic reviews and meta-analyses, for example, on sexual and reproductive health (Arthur et al., 2013); child mortality in relation to mother survival (Nguyen et al., 2019); HIV incidence in adolescents and young women in Africa (Birdthistle et al., 2019); child stunting and school performance (Gansaonré et al., 2022); HIV incidence in pregnancy and breastfeeding (Graybill et al., 2020); and migration and HIV acquisition (Dzomba et al., 2019). For these reasons it is important that the

complexities of HDSS data and analyses and reported appropriately, assuming that readers may not be aware of the platform-specific issues. It was noted in the meta-analysis on HIV incidence mentioned above, that sometimes the potential impact of migration was not totally clear in some HDSS analyses, making interpretation challenging (Birdthistle et al., 2019).

The objectives of this paper are to demonstrate the utility of using HDSS data for secondary analyses, and to provide guidance on the conduct and reporting of these analyses using examples from the literature. Firstly, examples of HDSS analyses which leverage the key aspects of HDSSs, the longitudinal nature and the linkages between individuals, are identified through a systematic search of the literature. These papers are used as the basis to describe 6 key aspects related to the conduct and reporting of HDSS analyses. Finally recommendations to researchers using HDSS data are given. It is anticipated that this review will be useful to producers of HDSS data, and new and existing users of the data. While the focus of the review is on ‘complex’ HDSS analyses, many of the concepts described apply to all uses of HDSS data.

4.3. Methodology

To identify a broad range of HDSS analyses to review, a systematic search of the literature was conducted using Google Scholar, as this tool tends to find literature from many different disciplines, and also covers the ‘grey literature’: conference papers, working papers and student theses. The search was restricted to papers published from 2012 up to the end of 2022. The search terms were based on the types of analyses known to utilise the longitudinal and linkage aspects of HDSSs:

("Demographic Surveillance System" OR "Demographic Surveillance Site") AND ("event history analysis" OR "joint modelling" OR "latent transition analysis" OR "lifecourse analysis" OR "life course analysis" OR "longitudinal analysis" OR "multi-state modelling" OR "sequence analysis" OR "seroconversion" OR "sero-conversion" OR "survival analysis" OR "trajectory analysis" OR "typology")

The tool ‘Publish or Perish’ was used to output the search results into a csv file; a total of 1745. The titles, abstracts or full text were scanned to identify duplicates, papers that were not in English, were BSc. or MSc. projects or were not analytical (i.e. protocols, review papers etc.) and analyses that were not involving humans, were laboratory or qualitative

studies, or were quantitative studies but not carried out in HDSSs: using these criteria, 1250 results were excluded (table 4.1).

The criteria for inclusion were firstly that the analysis must be using the HDSS data longitudinally and secondly that it must use some level of data manipulation which leverages linkages within each individual's HDSS repeated surveys, and/or linkages between individuals within households or family groups. Thus, papers which just analysed a cross-sectional survey, or a standard mortality rate analysis, were not included. Analyses of studies that solely used the HDSS as a platform, i.e. collected data from HDSS members but did not link to other HDSS data or other nested studies, were also excluded, unless there was at least 3 linked data points. Using these criteria a further 454 results were excluded (table 4.2).

Table 4.1: records excluded from systematic review as not HDSS analyses

Reason	N
Duplicates	84
Not English	36
BSc/MSc thesis	60
Non analytical (book chapter, protocol, review etc.)	346
Laboratory	177
Not human	9
Qualitative	53
Non-HDSS (DHS or other survey, hospital study etc.)	485
Total	1250

Table 4.2: reasons HDSS records excluded from systematic review

Reason rejected	N
Not longitudinal	152
New data collected to compare to DSS results	4
Intense primary data collection	2
Followed-up migrants outside	10
HDSS as platform but not really linked to DSS data	113
"Standard" longitudinal analyses	170
Not enough information to use	3
Total	454

A total of 41 papers were included from the systematic review; a further 5 were added following additional searches undertaken to identify papers where the term ‘demographic surveillance site’ was not used, in particular to search for analyses of HDSSs which were not represented on the list of included papers. It is acknowledged that the search was not exhaustive, however the coverage of HDSSs, topics and techniques was felt to be broad enough to conduct the review.

The 46 papers used data from 14 HDSSs (NB. 7 use data from multiple HDSSs some of which may not be listed below), most HDSSs only contribute one paper, uMkhanyakude has the largest number with 12, Agincourt 7, Nairobi 6, Karonga 3 and Rufiji 2. Most of the HDSSs which contribute papers are in sub-Saharan Africa, but Thailand and Bangladesh are also represented (table 4.3).

Table 4.3: papers selected for in-depth review by HDSS

HDSS name	Other names	Country	HDSS start year	Number of papers
Agincourt		South Africa	1992	7
Cuatro Santos		Nicaragua	2004	1
Farafenni		Gambia	1981	1
Kanchanaburi		Thailand	2000	1
Karonga		Malawi	2002	3
Kyamulibwa	Masaka, Kalungu	Uganda	1989	1
Magu	Kisesa	Tanzania	1994	1
Manicaland		Zimbabwe	1996	1
Matlab		Bangladesh	1966	1
Nairobi		Kenya	2002	6
Niakhar		Senegal	1962	1
Rakai		Uganda	1994	1
Rufiji		Tanzania	1998	2
Umkhanyakude	Hlabisa, Africa Centre, AHRI	South Africa	2000	12
Multiple HDSS				7
Total				46

There were several examples of data reduction techniques and complex data approaches being used which were not include as the data were not used longitudinally, either just using one round of HDSS data or using the latest data available for each participant: for example

in Agincourt HDSS Bayesian belief network models were used to examine household food security (Eyre et al., 2021) and ‘mixture of factor analyzers for mixed data’ models used to assign households to clusters based on socio-economic factors (McParland et al., 2014). In Cuatro Santos HDSS in Nicaragua cluster analysis was used to assign households to groups based on multi-dimensional poverty (Källestål et al., 2020). There were also examples of usage of data linkage within households to generate household composition variables, however the overall analysis was not longitudinal, so they were not included (as any one-off household survey could have done the same thing), for example an analysis from the Mekong HDSS in Cambodia which looked at living arrangements and the association with school attendance (Heuveline and Hong, 2017), and a similar analysis from Agincourt in South Africa (Madhavan et al., 2017b). There were also some examples of data science techniques being used on verbal autopsy data regarding improving cause of death algorithms (Murtaza et al., 2018) and several examples where spatial clustering was assessed, sometimes longitudinally (Becher et al., 2016). These were not included as they do not really harness the individual HDSS linkages. Several HDSSs have carried out multiple rounds of HIV testing enabling estimates of HIV incidence, these papers were only included if they harnessed another aspect of HDSS linkage, i.e. linkage between couples or household members, or linkage to another HDSS outcome such as migration or mortality.

The in-depth review of each selected paper focused on the methods and results sections, plus the limitations sections of the discussion to assess them according to 6 key aspects of using HDSS data:

1. Data manipulation methods;
2. Dataset structure;
3. Statistical methods;
4. How repeated measures were used or accounted for;
5. How in/out-migration was dealt with; and
6. How other missing data were dealt with.

4.4. Results

46 papers were included for in-depth review; the subjects under analysis included mortality, fertility, migration, household composition, adolescent transitions, HIV diagnosis and care, childhood growth plus one paper analysing participation in surveys. All papers are listed in table 4.4 and may be referred to below in the summary of findings.

4.4.1. Example analyses

Several papers of particular interest are listed below to demonstrate the breadth of topics and techniques used on HDSS datasets:

1. This example used many secondary sources to answer a novel question on HIV sero-conversion: data from uMkhanyakude HDSS (South Africa) were used to create a time-to-event dataset from the date the participant was first known to be HIV negative ending at the estimated date of sero-conversion, censor dates were added from the HDSS residency dataset (dates of migration/death) and times of potential exposure identified through an annual survey where sexual partners were reported (Harling et al., 2014).

2. This analysis manipulated data to construct a dataset with the HIV/ART status for each day of the 1000-day period from estimated conception of a live-born child to estimated end of breast-feeding, and used sequence analysis and cluster analysis to assess different patterns of engagement with care. Data from Agincourt HDSS (South Africa) were used to identified women with a live birth and linked clinic data to identify date of HIV sero-conversion, date of ART initiation and any time they were classed as disengaged from care. Additionally time to event datasets were created starting at conception date and ending at seroconversion or starting ART and these events assessed using survival analysis (Etoori et al., 2021).

3. This example demonstrated the value in pooling HDSS data as their large dataset enabled assessment of child mortality risk at quite specific times around mother's death or migration, and pregnancy of subsequent sibling, i.e. 6-3 months before mother’s death, 3 months-15 days before mother’s death, from 15 days before to 15 days after mother’s death etc. Data from 29 HDSSs were used, linking data from children aged under 5 with their mothers and siblings’ dates of birth, migration and death which were used to create time-varying variables on a time to event dataset with the outcome of child mortality. (Bocquier et al., 2021).

4.4.2. Data manipulation techniques

Data manipulation techniques were classified first into those which took advantage of linking data from multiple time-points from one individual, this was used in 5 different ways:

1. to create summary exposure or outcome measures which were not linked to a date/time, i.e. child anthropometry data from multiple time points from Kyamulibwa HDSS in Uganda

were used to create a summary variable to indicate child's experience with stunting (i.e. always stunted, never stunted, improved, worsened), this variable was used as an exposure for a later analysis (Asiki et al., 2019);
2. to create an event with a date as an end-point i.e. when multiple rounds of HIV test and clinic data from uMkhanyakude HDSS in South Africa were linked to generate dates of specific transitions along the HIV care pathway which were used in multiple time-to-event analyses (Haber et al., 2017);
3. to create time-varying exposure variables i.e. when data from multiple sexual behaviour surveys from uMkhanyakude HDSS in South Africa were used to create time-varying exposures relating to age of sexual partners in a time-to-event analysis with HIV sero-conversion as the outcome (Harling et al., 2014);
4. a sequence of events/states i.e. in Agincourt HDSS in South Africa, HIV test, clinic and treatment data on women with a live birth were combined to generate a sequence with one variable per day from estimated date of conception to estimated date of cessation of breast-feeding (Etoori et al., 2021); or
5. Lagged data used as exposure, i.e. in uMkhanyakude HDSS in South Africa, data collected at age 11-13 was used as exposure for outcomes in teen mothers at age 20 (Ardington et al., 2015).

The second main data manipulation technique involved linking data between individuals from 1 or many time points, there were also 5 separate techniques:

1. Linkage for matched analysis i.e. data from maternal sisters (linked through mother id) were matched for an analysis of teen pregnancy in uMkhanyakude HDSS in South Africa (Ardington et al., 2015);
2. Data from linked person(s) used to create exposure variable for index i.e. an analysis using pooled data from multiple HDSSs used linkage between index children and their mothers and siblings to create time-varying exposure variable relating to mother and sibling survival and migration status (Bocquier et al., 2021);
3. Data from linked household members used to create summary exposure or outcome variables i.e. in Nairobi HDSS in Kenya, data from household members' education level were used to create household summary variable used as an exposure variable (Mutisya et al., 2016);
4. As 3 but the linked unit was something other than the household i.e. data from Cuatro Santos HDSS in Nicaragua summarised data from adolescent girls in small geographical areas to generate background adolescent pregnancy rates to use in models (Pérez et al., 2021);

5. Data from linked household members used to create household event with date i.e. individuals' death and migration data from Rakai HDSS in Uganda were used to generate household events of dissolution or migration (Muniina, 2016).

4.4.3. Dataset structures

The resulting datasets were commonly 'time-to-event' with an end point outcome under analysis: this is a very common data structure when analysing longitudinal HDSS data. The majority of these used continuous episodic data, but some simplified the dataset by splitting the data one record per month (Houle et al., 2021) or year (Madhavan et al., 2012) that each participant was present. Others reduced the multiple records to one record per person (Machiyama et al., 2015) and most of the remainder ended up with a multi-record dataset which were not typical time-to-event set-ups (Schatz et al., 2015). Most datasets used individuals as the unit of analysis, but there were a few that used the household (Sartorius et al., 2014). One analysis reduced the data into contingency tables with 1 record per combination of all included variables (Dobra et al., 2019).

4.4.4. Statistical methods

The majority of statistical techniques used were fairly standard for epidemiology and demography data, survival (time-to-event) analyses (including Kaplan Meier figures, Cox regression, Poisson regression), sometimes allowing for competing events (Haber et al., 2017; Kim et al., 2020), and one using landmark analyses (repeating the model only including people with data from certain ages onwards) (Sunny et al., 2019), and logistic or linear regression which may be multinomial (Korinek and Punpuing, 2012). The analyses which had a time-to-event dataset split into monthly records, rather than episodes, used multi-level modelling to allow for the data structure (Oketch et al., 2012). There were a couple of example of sequence analysis (Etoori et al., 2021; Larmarange et al., 2015), one that used Sankey diagrams to display descriptive data (Larmarange et al., 2015), two using multi-state transition models (Bagayoko et al., 2020; Oduro et al., 2022), one conditional inference tree analysis (Pérez et al., 2021), and a few that used Bayesian models (Dobra et al., 2019; Risher et al., 2021; Sartorius et al., 2014). Only one analysis used joint modelling (Risher et al., 2021). Joint modelling is now often preferred over more traditional techniques, such as Cox regression, in clinical trial survival analyses as it allows for more accurate and precise estimated of treatments effects (Gould et al., 2015). It can account for informative participant drop-out (see section on migration below) and techniques are being developed that allow for left (Crowther et al., 2016) and interval (Lovblom et al., 2023) censoring, making it likely that it will become a useful technique for analyses of HDSS data.

4.4.5. Repeated measures

There are a few issues regarding repeated measures in HDSS data: inconsistencies in data repeatedly collected (i.e. age at first marriage reported differently at different time points), repeated events/outcomes either within one individual (i.e. school drop-out may happen multiple times) or within another binding entity (i.e. one mother may have more than one child in the analysis), or data formatting resulting in multiple records per unit of analysis (i.e. episodic data changed to 1 record per person per month). The studies in the review dealt with data inconsistencies in different ways: a study on sexual debut and schooling using data from Karonga HDSS used the youngest reported age of sexual debut if there were inconsistencies (Sunny et al., 2019), however a study in Manicaland HDSS in Zimbabwe also looking at sexual behaviour data excluded participants if inconsistencies could not be resolved even though it reduced their sample size from 28,073 to 11,647 (Del Fava et al., 2016); a few did not mention any issue with inconsistencies though it seems likely that there were some, i.e. a study in uMkhanyakude HDSS in South Africa used data from several sources to generate an indicator of teen pregnancy but did not mention how they prioritised the data sources (Ardington et al., 2015). In the studies which used reported dates to identify transitions, the earliest date tended to be used and often the assumption was made that the person could only experience the event or transition once, i.e. in Rufiji HDSS in Tanzania a time-to-event analysis of predictors of parental absence only used the first instance (Gaydosh, 2015). The majority of analyses accounted for repeated records or clustering in the statistical methods used, either by introducing fixed or random effects, or using a method that intrinsically accounts for multiple records, i.e. survival analysis or multi-level modelling.

4.4.6. Migrations

HDSSs only record data on participants if they are living within the area, this can make linked analyses challenging as only using data from people who stay in the area may introduce bias. In ‘typical’ time-to-event analyses which use HDSS data, participants contribute time ‘at risk’ when they are present, i.e. no-one is excluded for not being present the whole time. This may be ‘right’ or ‘left’ censoring if a participant out-migrates or in-migrates, or ‘interval’ censoring, if they out-migrate and then return. Many of the papers in the review used this approach, some however restricted their analysis to children who were born in the area (usually because exposure data were not available for those who in-migrated) but did not exclude those who out-migrated before experiencing the event, for example a study using data from Farafenni HDSS in the Gambia used a time-to-event analysis starting from birth of children in the area and children were censored if they or their mother left the area (as the exposure was mother’s vital status) (Scott et al., 2017). Other

studies used migration as an outcome or exposure (Muniina, 2016; Ziraba, 2013) so bias was reduced. Some of the studies using longitudinally linked data excluded all participants who did not have all data points, for example a study using data from Nairobi HDSS in Kenya used anthropometry data during childhood linked to a survey in adolescence where only 692 of 3419 children had all the data required (Oduro et al., 2022), or a study from Matlab HDSS, Bangladesh which linked fertility intention data from 1990 with data on linked children born up to 7 years later and their survival up to 3 years (Bishai et al., 2015). Other studies tried to mitigate the effect of migration by a. trying to keep the inclusion criteria as wide as possible, i.e. only requiring data from 2 time-points (Harling et al., 2014), or b. treating each transition/event as a separate analysis so as many participants can be included in at least one (Haber et al., 2017), or c. include time spent in the HDSS as a control measure (Fotso et al., 2013), d. triangulate the findings with a slightly different analytical approach, i.e. one study carried out a separate matched analysis of sisters to add to their findings as they felt it may have been affected by attrition (Ardington et al., 2015).

Surprisingly many papers made no mention of migration or attrition in the discussion, even when the analysis design makes it seem like migration might have caused some bias. For example, a study using data from Cuatro Santos HDSS in Nicaragua included women aged 20-24 who were present in the HDSS in the 2014 survey (using data from previous studies/data to assign teen pregnancy status) so women experiencing teen pregnancy who left the area would not be included. However there was no mention of this in the discussion (Pérez et al., 2021). Others did discuss the issues: i.e. 1. a study using data from Kyamulibwa HDSS in Uganda needed participants to have data from 4 consecutive surveys, they recognised this as a potential limitation but also reported that 79% of those in the 1st survey had data from all 4, so felt that the data could still be representative (Asiki et al., 2019); 2. a study of changes in household composition in older people using data from Agincourt HDSS, South Africa recognised that not being able to include recent in-migrants may have affected the results as they might be different to those included (Madhavan et al., 2017c); and 3. a study in Kanchanaburi HDSS in Thailand did a time-to-event analysis with school drop-out as the exposure, they did not include in-migrants as they were using exposure data from a specific survey, they recognised this as a potential issue though reasoned that as they only had a relatively short follow-up time there was not likely to have been a big effect (Korinek and Punpuing, 2012). A few attempted to examine the effect of attrition, by comparing attributes of those included and not (Sunny et al., 2018), or assessing the effect of the exposure on the outcome of out-migration as well as the outcome under investigation (Ardington et al., 2014; Finlay et al., 2015).

4.4.7. Missing data

The final aspect of information that extracted during the review of the papers was on missing data. All studies, regardless of whether the data come from an HDSS or not, may be prone to issues from missing data and the majority of studies either dealt with missing data in standard ways, i.e. excluding those missing data from the model (Sunny et al., 2018), retaining them under a category of unknown/missing (Ginsburg et al., 2016), with a few attempting some multiple imputation (Anekwe et al., 2015; Harling et al., 2014; Marston et al., 2013). There were a few instances of data from previous or later time points being used to impute missing data, though no discussion over the pros and cons of this approach (Bagayoko et al., 2020; Gaydosh, 2017; Nyirenda, 2014). In the analyses which utilised linkage between participants (i.e. mother-child, or within households) to generate exposure or outcome variables, there was never any discussion of what was done if there were any missing data when creating variables i.e. in an analysis using data from Nairobi HDSS in Kenya the average number of years of schooling per household was calculated, but there was no mention of what was done if any of the data were missing (Mutisya et al., 2016).

4.5. Discussion

4.5.1. Conclusions of literature review

46 published papers using complex data manipulations on longitudinal HDSS data were identified, from only 14 HDSSs, with some particular sites over-represented. There were 60 HDSSs identified through the literature search, however some may not be ‘full’ HDSSs, i.e. they may only include certain age groups. The sites with the largest number of papers included in the review, uMkhanyakude, Agincourt, and Nairobi, also had the highest number of other papers identified, but excluded, from the systematic review. These 3 sites have been running for a long time, though are by no means among the ‘long-runners’ among HDSSs. Several HDSSs had a good number of other papers, including Matlab in Bangladesh, Nouna in Burkina Faso, and KEMRI and Kilifi, both in Kenya. Further analyses, including complex ones, may be possible with data from these and other HDSSs. Many of the papers reviewed are very high quality and describe some really interesting work. There was a high number of data manipulation techniques used which really demonstrates the flexibility of HDSS data. Some quite complex statistical techniques were used, however the majority usage of more ‘standard’ techniques show that advanced statistical skills are not a requirement for conducting valuable analyses with HDSS data. While there were some interesting and novel ways used to approach the issues of repeated data, migration and missing data it was most striking how often they were not discussed.

4.5.2. Recommendations to researchers

Producers of HDSS data should assess whether their data are being used to the fullest; whether through more complex analyses as listed in this paper, or more standard analyses. The sections on data manipulation, dataset structures and statistical methods, plus the listing of example analyses may help to inspire further analyses.

Users of HDSS data should be aware of the issues of repeated data capture, of in- or out-migration, and of missing data, and consider the most appropriate way of dealing with it. This may require close consultation with HDSS data producers.

Researchers writing up HDSS analyses for publication should consider whether their approach to dealing with repeated data, migrations and missing data have been appropriately detailed in the methods section. Equally, researchers should ensure to include adequate discussion of these factors in the limitations section, to enable readers who are not experts in HDSS data to fully understand any potential issues in interpretation of findings.

Table 4.4: Listing of all HDSS analyses reviewed, coded according to the 6 aspects (see footnote for meanings)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Missing data	Reference
The causal effect of childhood measles vaccination on educational attainment: a mother fixed-effects study in rural South Africa	uMkhanyakude, South Africa	1995-2007	Measles vaccination and education	A; B	D	A	D; E	F	D	(Anekwe et al., 2015)
The economic consequences of AIDS mortality in South Africa	uMkhanyakude, South Africa	2000-2009	Impact of mortality of household SES	I	E	A	D; E	A	C	(Ardington et al., 2014)
Early childbearing, human capital attainment, and mortality risk: Evidence from a longitudinal demographic surveillance area in rural KwaZulu-Natal, South Africa	Umkhanyakude, South Africa	2001-2012	Teen pregnancy and association with schooling and mortality	A; C; E; F	D	A	A; D; E	E	C	(Ardington et al., 2015)
The effect of childhood stunting and wasting on adolescent cardiovascular diseases risk and educational achievement in rural Uganda: a retrospective cohort study	Kyamulibwa , Uganda	2001-2011	Change in stunting status in childhood used as exposure for NCD risk factor and education status.	A	D	A	D	D	A	(Asiki et al., 2019)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Missing data	Reference
Understanding wealth transitions among households in urban slums of Nairobi: A multi-state transition modelling approach	Nairobi, Kenya	2005-2015	Household wealth transitions	B	C	F	D; E	A	B; E	(Bagayoko et al., 2020)
Selection bias in the link between child wantedness and child survival: theory and data from Matlab, Bangladesh	Matlab, Bangladesh	1990-2000	Child mortality in relation to their mother's pre-conception report of desire for more children	J	A	A; B	D	D	A	(Bishai et al., 2015)
The Crucial Role of Mothers and Siblings in Child Survival: Evidence From 29 Health and Demographic Surveillance Systems in Sub-Saharan Africa	Multiple, Africa	1990-2016	Effects of mother & sib presence, death, migration and birth intervals on child mortality	J	A	B	D	A	A	(Bocquier et al., 2021)
Improving the validity of mathematical models for HIV elimination by incorporating empirical estimates of progression through the HIV treatment cascade	Umkhanyakude, South Africa	2004-2014	Multiple transitions along HIV care pathway assessed for their effect on life expectancy &	B	A	B	A	E	A	(Chang et al., 2018)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Missing data	Reference
			potential onward transmission							
Young children's probability of dying before and after their mother's death: a rural South African population-based surveillance study	Agincourt, South Africa	1994-2008	Child mortality in relation to timing and cause of mother's death	J	B	C	D	A	A	(Clark et al., 2013)
HIV-seroconversion among HIV-1 serodiscordant married couples in Tanzania: a cohort study	Magu, Tanzania	2006-2016	Factors associated with HIV sero-conversion in HIV discordant couples	J	A	B	E	A	A	(Colombe et al., 2019)
Transition to Parenthood and HIV Infection in Rural Zimbabwe	Manicaland, Mozambique	1998-2011	Sequences from sexual debut, union formation and child-bearing created and grouped and checked for association with HIV	A; C	D	A	C	E	D	(Del Fava et al., 2016)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Miss-ing data	Reference
			status and birth cohort							
Space-time migration patterns and risk of HIV acquisition in rural South Africa	uMkhanyakude, South Africa	2000-2014	Migration & HIV acquisition	A; B	A	B	E	C; F	B	(Dobra et al., 2017)
A method for statistical analysis of repeated residential movements to link human mobility and HIV acquisition	uMkhanyakude, South Africa	2004-2016	Migration & HIV acquisition	B	F	H	E	C	B	(Dobra et al., 2019)
Patterns of engagement in HIV care during pregnancy and breastfeeding: findings from a cohort study in North-Eastern South Africa	Agincourt, South Africa	2014-2018	HIV diagnosis and treatment through pregnancy and breast-feeding period examined	B; C	A; D	A; B; E	E	B; C	C	(Etoori et al., 2021)
The effects of maternal mortality on infant and child survival in rural Tanzania: a cohort study	Multiple, Tanzania	1996-2012	Maternal mortality and subsequent child mortality	J	A	B	D	A	C	(Finlay et al., 2015)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Missing data	Reference
Birth spacing and child mortality: an analysis of prospective data from the Nairobi urban health and demographic surveillance system	Nairobi, Kenya	2003-2009	Birth spacing and child mortality	G; J	A	B	D	B; E	D	(Fotso et al., 2013)
Childhood risk of parental absence in Tanzania	Rufiji, Tanzania	2001-2011	Risk of parental absence for any cause from birth to age 10	J	C	B	A	B; C	B; E	(Gaydosh, 2015)
Beyond Orphanhood: Parental Nonresidence and Child Well-being in Tanzania	Rufiji, Tanzania	1998-2011	Effect of parental non-residence on child mortality and entry into school at appropriate age	J	A	B	D	B; C	B; E	(Gaydosh, 2017)
Healthy or Unhealthy Migrants? Identifying Internal Migration Effects on Mortality in Africa using Health and Demographic Surveillance Systems of the INDEPTH …	Multiple, Africa	1998-2012	Migration status categorised as never, in-migrant and returned migrant and used as exposure for mortality risk in adults	D	A	B	D	C	C	(Ginsburg et al., 2016)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Missing data	Reference
From HIV infection to therapeutic response: a population-based longitudinal HIV cascade-of-care study in KwaZulu-Natal, South Africa	umkhanyakude, South Africa	2006-2011	Speed of multiple transitions along HIV care pathway	B	A	B	B	E	B	(Haber et al., 2017)
Do age-disparate relationships drive HIV incidence in young women? Evidence from a population cohort in Rural KwaZulu-Natal, South Africa	umkhanyakude, South Africa	2003-2012	HIV sero-conversion as outcome, age disparity with sexual partner is exposure	B; D	A	B	D	E	D	(Harling et al., 2014)
Household context and child mortality in rural South Africa: the effects of birth spacing, shared mortality, household composition and socio-economic status	Agincourt, South Africa	1994-2008	Child mortality & household composition	G; J	B	C	D	A	B	(Houle et al., 2013)
Linking the timing of a mother's and child's death: Comparative evidence from two rural South African population-based surveillance studies, 2000–2015	Multiple, South Africa	2000-2015	Child mortality, association with maternal mortality and presence of relatives	G; J	B	C	D	B	A	(Houle et al., 2021)
HIV seroconcordance among heterosexual couples in rural KwaZulu-	uMkhanyakude, South Africa	2003-2016	HIV incidence & concordance	B; H; J	A	B	E	A	E	(Kim et al., 2020)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Missing data	Reference
Natal, South Africa: a population-based analysis										
The effect of household and community on school attrition: An analysis of Thai youth	Kanchanaburi, Thailand	2001-2004	Household composition and school drop-out in adolescents	G; H	C	A	D; E	B; C	C	(Korinek and Punpuing, 2012)
Participation Dynamics in Population-Based Longitudinal HIV Surveillance in Rural South Africa	umkhanyakude, South Africa	2003-2012	Participation in repeated HIV survey rounds examined for patterns	C	D	A; E; I	E; F	E	A	(Larmarange et al., 2015)
An assessment of childbearing preferences in northern Malawi	Karonga, Malawi	2008-2015	Fertility intentions and subsequent birhts	J	D	A	E	A	C	(Machiyama et al., 2015)
Child mobility, maternal status, and household composition in rural South Africa	Agincourt, South Africa	1999-2008	Household level variables and their association with child migration	G; J	C	C	D	A	B	(Madhavan et al., 2012)
Social positioning of older persons in rural South Africa: change or stability?	Agincourt, South Africa	2000-2010	Change in household composition in older people	A	E	D	D	B; D	A	(Madhavan et al., 2017c)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Miss-ing data	Reference
Is the risk of HIV acquisition increased during and immediately after pregnancy? A secondary analysis of pooled HIV community-based studies from the ALPHA network	Multiple	1989-2012	HIV incidence in relation to pregnancy	B; J	A	B	E	A; F	D	(Marston et al., 2013)
Household survival and changes in characteristics of households in rural South-western Uganda through the period of 1989 to 2008 [Phd thesis]	Rakai, Uganda	1989-2008	Household level variables were created (composition, presence of person with HIV) and used to test association with household dissolution, migration or change in membership	G; I	C	D	E	A; C	A	(Muniina, 2016)
The effect of education on household food security in two informal urban settlements in Kenya: a longitudinal analysis	Nairobi, Kenya	2007-2012	Summary household measures used to assess association with food security	G	E	A	D	A	A	(Mutisya et al., 2016)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Miss-ing data	Reference
How do migrations affect under-five mortality in rural areas? Evidence from Niakhar, Senegal	Niakhar, Senegal	1998-2013	Summary data on migration of household members tested for association with occurrence of U5 mortality in household	A; E; G	E	A	D	C; F	A	(Nguemdjo and Ventelou, 2021)
Ageing with HIV: An investigation of the health and well-being of older people in a rural South African population with a severe HIV epidemic [Phd thesis]	umkhanyakude, South Africa	2005-2010	Changes in living arrangements for older people	A; G	D; E	A	E	D	C; E	(Nyirenda, 2014)
A multi-state transition model for child stunting in two urban slum settlements of Nairobi: a longitudinal analysis, 2011-2014.	Nairobi, Kenya	2010-2014	Transition between stunting categories in young children	B	D	F	E	D	A	(Oduro et al., 2022)
Do poverty dynamics explain the shift to an informal private schooling system in the wake of free public primary education in Nairobi slums?	Nairobi, Kenya	2005-2009	Exposure is change in HH SES and outcome is transfer from public to private school	A	E	C	D	E	C	(Oketch et al., 2012)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Miss-ing data	Reference
Trends and factors related to adolescent pregnancies: an incidence trend and conditional inference trees analysis of northern Nicaragua demographic …	Cuatro Santos, Nicaragua	2001-2014	Adolescent child-bearing and associations with living arrangements and other household factors	A; G; H	D	G	F	F	A	(Pérez et al., 2021)
Age patterns of HIV incidence in eastern and southern Africa: a modelling analysis of observational population-based cohort studies	Multiple (ALPHA)	1989-2017	HIV incidence	B	A	H	D; E	A; E	B	(Risher et al., 2021)
The dynamics of household dissolution and change in socio-economic position: A survival model in a rural South Africa	Agincourt, South Africa	1993-2008	Summary household measures used to assess association with household dissolution	G; I	A	H	D	A	C	(Sartorius et al., 2014)
Dependent or productive? A new approach to understanding the social positioning of older South Africans through living arrangements	Agincourt, South Africa	2000-2010	Household composition in older people	G	E	D	D	A	A	(Schatz et al., 2015)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Missing data	Reference
Effect of maternal death on child survival in rural West Africa: 25 years of prospective surveillance data in The Gambia	Farafenni, Gambia	1989-2014	Maternal vital status and child survival	J	A	B	D	B	A	(Scott et al., 2017)
An intimate epidemic: HIV and marriage in rural Uganda [Phd thesis]	Masaka & Rakai, Uganda		HIV incidence & concordance	B; J	A	B	E	A; F	B; E	(Sully, 2015)
Does early linear growth failure influence later school performance? A cohort study in Karonga district, northern Malawi	Karonga, Malawi	2002-2015	Changes in stunting in childhood used as exposure for later school performance (age for grade)	A	D	A	F	D	B	(Sunny et al., 2018)
Lusting, learning and lasting in school: sexual debut, school performance and dropout among adolescents in primary schools in Karonga district, northern Malawi	Karonga, Malawi	2007-2016	Age at sexual debut examined for association with school drop-out	B; D	A	B	A; D	A	A	(Sunny et al., 2019)
Use of antiretroviral therapy in households and risk of HIV acquisition in rural KwaZulu-Natal, South Africa, 2004–12: a prospective cohort study	uMkhanyakude, South Africa	2004-2012	Household ART usage and HIV acquisition	B; J	A	B	D; E	A	A; C	(Vandormael et al., 2014)

Title	HDSS, country	period	Subject	Data manipulation	Data-set	Statistics	Repeated data	Migration	Miss-ing data	Reference
Adult mortality and its impact on children in two informal settlements in Nairobi, Kenya [PhD thesis]	Nairobi, Kenya	2003-2007	Child migration following death in household	J	C	A	D	A	C	(Ziraba, 2013)

Codes:
Data manipulation: A=Data from >=2 time points used to create summary exposure or outcome measure; B=Data from >=2 time points used to create 'event' with date; C=Data from >=2 time points used to create sequence; D=Data from >=2 time points used to create time-varying exposure; E=Lagged data used as exposure; F=Matched analysis; G=Data from linked household members used to create summary exposures or outcomes; H=Data from members of other groups used to create summary exposures or outcomes; I=Data from linked household members used to create household event with date; J=Data from linked person(s) used to create exposure for index
Dataset: A=Time to event (continuous); B=Time to event (monthly); C=Time to event (yearly); D=One record per person; E=Multi-record; F=Summaries
Statistical methods: A=Logistic/linear regression; B=Survival analysis; C=Multi-level modelling; D=Descriptive only; E=Sequence analysis; F=Multi-state transition model; G=Conditional inference tree analysis; H=Bayesian modelling; I=Sankey diagrams
Repeated data: A=First event/report used; B=Events treated separately; C=Excluded inconsistencies; D=Accounted for in model; E=Unclear; F=N/A
Migration: A=Contribute time when present; B=In-migrants excluded; C=Part of analysis; D=All data points needed; E=Attempt to reduce effect; F=Potential effect of out-migration
Missing data: A=Not mentioned/none; B=Excluded (cases or variables); C=Missing kept as category; D=Statistical imputation; E=Imputed from other data points

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	145184	**Title**	Ms
First Name(s)	Estelle		
Surname/Family Name	McLean		
Thesis Title	Demonstrating the value of Health and Demographic Surveillance Site data for complex secondary analyses, illustrated with analyses of young people's living arrangements and transitions to adulthood.		
Primary Supervisor	Rebecca Sear		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Wellcome Open Research
Please list the paper's authors in the intended authorship order:	Estelle McLean, Albert Dube, Fredrick Kalobekamo, Emma Slaymaker, Amelia C Crampin, Rebecca Sear
Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I developed the concepts for and created the datasets, and drafted the manuscript

SECTION E

Student Signature	
Date	27 July 2023

Supervisor Signature	
Date	27 July 2023

5. Family network and household composition: a longitudinal dataset derived from the Karonga HDSS, in rural Malawi

Estelle McLean*[1,2], Albert Dube[1], Fredrick Kalobekamo[1], Emma Slaymaker[2], Amelia C Crampin[1,2], Rebecca Sear[2]
*author for correspondence
1. Malawi Epidemiology and Intervention Research Unit
2. London School of Hygiene and Tropical Medicine

5.1. 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 could be used and provide information on limitations which affect the types of analyses that could be carried out.

Keywords
Family; Relatives; Household; GPS; Longitudinal; Malawi;

5.2. 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 and 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 and 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 Unit (MEIRU), formerly known as the

Karonga Prevention Study. It has been running since 2002, but built upon research infrastructure which had been ongoing in the same area since 1979 (Ponnighaus et al., 1987). Early research in the area focussed 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 be able to generate family lineages. This practice has continued up to present day, allowing the generation of this rich dataset.

5.3. Materials and methods

5.3.1. Context

The Karonga Health and Demographic Surveillance Site (HDSS) was established in 2002 in the southern part of the Karonga district in northern Malawi (Crampin et al., 2012). The area is largely rural with one semi-urban trading town, several smaller market villages and one port on Lake Malawi. The majority of the population engage in subsistence farming or fishing. The main ethnic group living here are Tumbuka, who since the 19th century have followed patrilineal and patrilocal customs: women tend to move to their husband's village when they marry (Malawi Human Rights Commission, 2006). In the event of divorce or even paternal death, children that are old enough to be away from their mother may be required to live with their father's family (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.

5.3.2. Initial data

The HDSS covers an area of 150km^2 and by 2016 had over 40,000 people under surveillance. Births and deaths are captured monthly through a system of local 'key informants', while migrations are captured annually through visits to all households. Specific dates for each event are captured and therefore the data are arranged as episodes which may start with initial census, birth or in-migration and end with death or out-migration. Participants are given a unique identifier which they retain in all studies: if they move household they are linked back to this ID (even if they leave the area and then return). 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, their episode at the old household is ended and a new one begun. In the HDSS, a household is defined as a group of individuals, rather than a location, meaning that if the group move they would still be classed as the same household. Household membership is defined by the participants through guidance of trained fieldworkers: all household members must usually live in the dwelling/compound together and recognise 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 be living in all the co-wives' households; all other participants may only belong to one household. GPS coordinates are recorded for each household at the initial census, when the household is established or if it moves. House move or change in household membership may result in one household being 'dissolved' and other(s) established. As 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. The parents and spouse 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 have been gathered regularly.

5.3.3. Data processing

Raw data are currently stored in Microsoft Access databases, and were extracted into Stata format. All data processing to create this dataset described in this paper was carried out 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 to 2017. The residency episodes are first reduced to one record per person per period by taking a snapshot on the mid-point of the period. This is to allow for more flexible data manipulation. As continuous data are available for all HDSS residents the length of the period represented by the snapshot can be varied according to the needs of the analysis (i.e. yearly, quarterly, monthly). This description will use the 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 listing 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 etc.), the family type: maternal (mother and any relatives through her [grand-parents, aunts/uncles, cousins etc.]);

paternal (father and any relatives through him), sister (half or full sister and any of her children or grand-children, brother (half or full brother and any of his children or grand-children), daughter (daughter and any her children or grand-children), son (son and any of his children or grand-children), the estimated genetic relatedness (i.e. 50% for parent-child, down to 3.125% for mother's cousin), categorised 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 step-cousins/aunts etc.), 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 the 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 are also available.

The population panel and the relationships dataset are used in 3 linked processes which generate variables describing the 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 3 processes are merged together so that all the above information is available for each person, at each time point they are present in the HDSS.

Household characteristics

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

The distance between each index household and every other household in the local area is then calculated. The summary household variables just described are 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 radiuses (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 for 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 2 or more households sometimes reside in very close proximity, sharing facilities in loose economic or social alliances, thus with shared resources and linked prospects. Using the population panel and relationships data, these grouped households were identified to generate an 'expanded household' definition. Grouped households are not formally identified during surveillance so a data-driven approach was used, harnessing the spacing between households at different population densities together with relationship data.

Initially, a random sample of 100 pairs of households 30m or less (but over 0m) apart were examined individually using satellite imaginary on Google Earth and assigned by eye as the same or different compounds. The 'same' households were a median of 7.7m (range 1.7-21.2m, IQR 4.1-11.4) apart while the 'different' ones were 18m (range 6-29.5m, IQR 13.7-21.3) apart. From this it was assumed that all households less than 5m apart were linked and may be linked if they were up to 20m apart. Individual households were assigned a 'starting' radius of 5, 10, 15 or 20 metres if the number of households within the radius was more than would be expected given the household density within 50m: a household in a more densely populated area would therefore have a smaller starting radius i.e. a household with 20 households within 50m ($7852m^2$) would expect to have 0.8 within 10m ($314.2m^2$) and 3.2 within 20m (1256.6m2) if they are distributed equally, so if they have 2 households within 10m and 3 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. Using the radius as a guide, households 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); not all households within the radius may be grouped if there is a bigger difference between them and the existing cluster. This method is prone to error, and but results in more appropriate connections between

households than using a more simple 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) household 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 household.

Family network

The GPS coordinates of all blood relatives (either singly or as groups i.e. maternal or paternal) are compared to those of the index for each time point. Summary variables are then calculated as either binary or continuous for presence of relatives within certain radiuses (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 the household relatives variables.

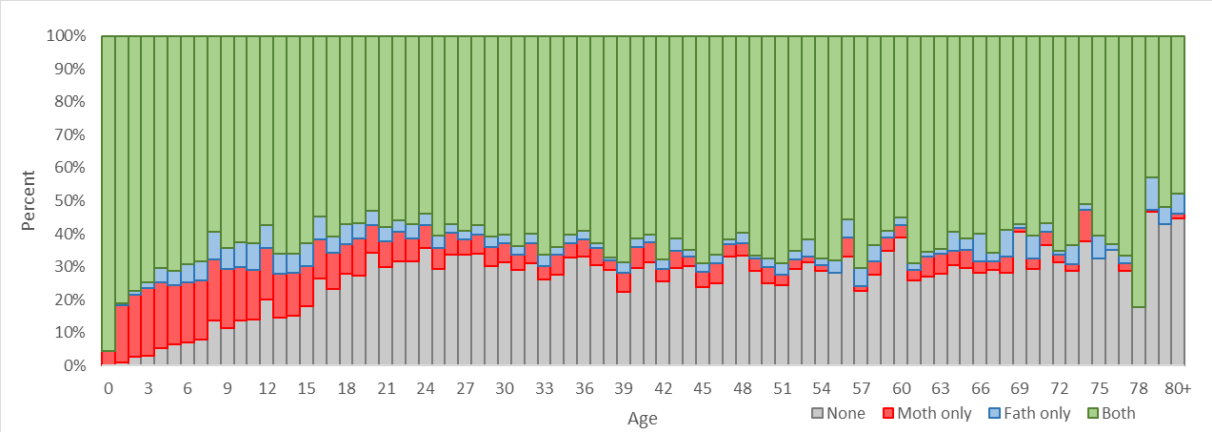
5.4. Examples of uses of dataset

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 the 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 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, in time 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 so may be able to generate similar datasets, following the logic described above.

5.5. Dataset validation / Limitations

While this dataset has many potential uses, it is important for users to be aware of some limitations to aid appropriate selection of data for analyses. The dataset is dependent on the parent and spouse links, which are not available for all HDSS members. The proportion of all HDSS members by single age year and whether their mother and father IDs are known is shown in figure 5.1. The proportion with at least one ID is highest for children, and there is 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.

Figure 5.1. percent of HDSS residents by age and availability of parent id-links



Being able to link individuals to their relatives also depends on whether other people have their parent/spouse id links. Figure 5.2 shows the average proportion of household relationships which are unknown, by the age of the index person and calendar year. Unsurprisingly the group with the lowest proportion of unknown relationships are children aged under 5, but the 30-49 year age group also have low levels (as their households are likely to be formed of their spouse and children). The groups with the highest proportion of unknown relationships are people aged over 70 and adolescents aged 15-19, however the proportions are not high (under 13%). By calendar year the proportions unknown decreased somewhat from 2004, but there was an increase at the very end of the period to 2017.

Figure 5.2. Average percent of relationships to index person within households which are unknown, by age group (of index) and year

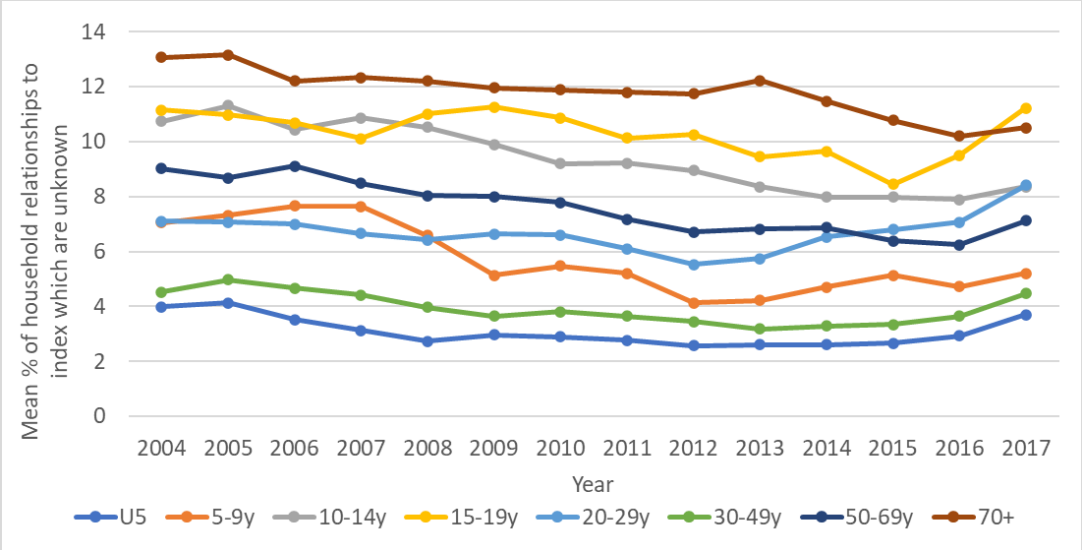


Table 5.1. Total number of individuals* in the dataset by age group, selected years and whether their relationship to other household members are fully known or fully unknown

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%

Age group	Households		2005	2007	2009	2011	2013	2015	2017
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%
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 that they are present in the HDSS

Table 5.2. Number of HDSS residents by age and sex, and how many years they were present

		Years present in the HDSS							
Sex & birth cohort		1-2y	3-4y	5-6y	7-8y	9-10y	11-12y	13-14y	Total
Male									
pre-1960	n	303	192	145	137	108	125	961	1,971
	%	15.4%	9.7%	7.4%	7.0%	5.5%	6.3%	48.8%	
1960-69	n	277	179	121	102	87	91	691	1,548
	%	17.9%	11.6%	7.8%	6.6%	5.6%	5.9%	44.6%	
1979-79	n	662	376	232	208	167	191	1119	2,955
	%	22.4%	12.7%	7.9%	7.0%	5.7%	6.5%	37.9%	
1980-89	n	1167	698	453	380	370	355	1263	4,686
	%	24.9%	14.9%	9.7%	8.1%	7.9%	7.6%	27.0%	
1990-99	n	1327	756	518	442	457	567	2459	6,526
	%	20.3%	11.6%	7.9%	6.8%	7.0%	8.7%	37.7%	
2000-9	n	1909	1024	769	914	1342	1291	2191	9,440
	%	20.2%	10.8%	8.1%	9.7%	14.2%	13.7%	23.2%	
post-2010	n	2201	1576	1145	739	0	0	0	5,661
	%	38.9%	27.8%	20.2%	13.1%	0.0%	0.0%	0.0%	
Total	n	7846	4801	3383	2922	2531	2620	8684	32,787
	%	23.9%	14.6%	10.3%	8.9%	7.7%	8.0%	26.5%	
Female									
pre-1960	n	328	263	170	145	162	167	1369	2,604
	%	12.6%	10.1%	6.5%	5.6%	6.2%	6.4%	52.6%	
1960-69	n	232	145	111	100	73	83	878	1,622
	%	14.3%	8.9%	6.8%	6.2%	4.5%	5.1%	54.1%	
1979-79	n	630	393	258	224	193	242	1269	3,209
	%	19.6%	12.2%	8.0%	7.0%	6.0%	7.5%	39.5%	
1980-89	n	1556	869	606	455	427	450	1317	5,680
	%	27.4%	15.3%	10.7%	8.0%	7.5%	7.9%	23.2%	
1990-99	n	2399	1328	978	769	695	751	1648	8,568
	%	28.0%	15.5%	11.4%	9.0%	8.1%	8.8%	19.2%	
2000-9	n	2352	1187	879	956	1473	1327	1787	9,961
	%	23.6%	11.9%	8.8%	9.6%	14.8%	13.3%	17.9%	
post-2010	n	2243	1591	1197	684	0	0	0	5,715
	%	39.2%	27.8%	20.9%	12.0%	0.0%	0.0%	0.0%	
Total	n	9740	5776	4199	3333	3023	3020	8268	37,359
	%	26.1%	15.5%	11.2%	8.9%	8.1%	8.1%	22.1%	

The actual number of individuals available in the dataset by year and age group are shown in Table 5.1, which also shows the proportion with complete information on their relationships to all household members and the proportion with no information at all. This shows that there are a high number of individuals with enough good data in all age groups, though the numbers do decrease after the age of 50.

The other potential limitation with the dataset is related to the HDSS data source: data are only available on participants when they are living in the HDSS area. Table 5.2 shows the number of HDSS residents by sex and birth cohort, and how many years they were present in the HDSS between 2004 and 2017 (maximum 14 years). While there are high numbers of participants who 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 to those who do not. Table 5.2 shows an effect of birth cohort (those with earlier birth dates are more likely to have complete data) and sex (males more likely to have complete data).

5.6. Dataset information

Ethics

The Karonga HDSS has ethical approval from the Malawi National Health Science Review Committee (approval #416) and from the London School of Hygiene and Tropical Medicine (approval #5081). All households provide written consent to take part in the Karonga HDSS, which may be rescinded at any time.

Data availability

Due to the detailed nature of the data describing exact living arrangement of participants, it is not possible to anonymise it sufficiently in a way that allows it to still be useful, thus the data are not available open access. However MEIRU welcomes requests to use the data from bona fide researchers, who should contact the first author (EM) in the first instance. Detailed documentation, including a listing of all available variables, can be found on the MEIRU data catalogue: http://kpsmw.lshtm.ac.uk/nada/index.php/catalog/13

Grant information

This work is supported by The Wellcome Trust (098610; 217073; through funds awarded to Amelia Crampin and The Karonga HDSS).

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	145184	**Title**	Ms
First Name(s)	Estelle		
Surname/Family Name	McLean		
Thesis Title	Demonstrating the value of Health and Demographic Surveillance Site data for complex secondary analyses, illustrated with analyses of young people's living arrangements and transitions to adulthood.		
Primary Supervisor	Rebecca Sear		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Journal of Biosocial Science
Please list the paper's authors in the intended authorship order:	Estelle McLean, Alison J Price, Luigi Palla, Emma Slaymaker, Amelia C Crampin, Albert Dube, Fredrick Kalobekamo, Rebecca Sear

Stage of publication	**Submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I developed the concept for the analysis, conducted all the data manipulations and analyses, wrote the manuscript and revised following comments from co-authors

SECTION E

Student Signature	
Date	27 July 2023

Supervisor Signature	
Date	27 July 2023

6. Data-driven versus traditional definitions of household membership and household composition: does latent class analysis produce meaningful groupings?

Estelle McLean[1,2#], Alison J Price[1,2], Luigi Palla[3], Emma Slaymaker[2], Amelia C Crampin[1,2], Albert Dube[1], Fredrick Kalobekamo[1], Rebecca Sear[2]

3. *Malawi Epidemiology and Intervention Research Unit, Lilongwe, Malawi.*
4. *London School of Hygiene and Tropical Medicine, Faculty of epidemiology and population health, London, United Kingdom*
5. *University of Rome La Sapienza, Rome, Italy*

 # author for correspondence: estelle.mclean@lshtm.ac.uk, ORCID: 0000-0001-6079-0663, London School of Hygiene and Tropical Medicine, department of epidemiology and population health, Keppel St, London WC1E 7HT, United Kingdom

6.1. Abstract

Adolescence is a key period of biological and social development and household living arrangements in adolescence in sub-Saharan Africa has been shown to be associated with multiple biosocial outcomes. Household is a commonly used term across a wide range of disciplines, however traditional, western-centric definitions are often used which may not capture important, context-specific differences in household membership (who belongs to which group) and composition (how the household members are related). This study used data on adolescents from rural northern Malawi from 2004-2016 to create context-relevant household composition variables using latent class analysis (LCA) with two household membership definitions, 'immediate' (as defined within the dataset) and 'expanded' (created to include relatives living close by). The extent to which different definitions of household composition alter observed associations with biosocial outcomes was investigated. LCA identified household compositions with greater complexity than those represented in western-centric definitions (LCA classes included 'brother's family', 'sister's family', 'maternal' and 'paternal'), with few individuals living in 'nuclear' families. Using the 'expanded' household definition created classes which, for example, distinguished between 'single mother' households and those with a single mother but living very close to maternal family. LCA was found to be most useful for guiding the creation of manual 'LCA-guided' variables to produce household composition definitions which were suitable for use as predictor variables. Compared to western-centric definitions, LCA-guided household

composition definitions using both ‘immediate’ and ‘expanded’ definitions provided greater detail about the contribution of household composition to variation in associations with biosocial outcomes: for example female adolescents in ‘maternal’ households had higher odds of a poor educational outcome, while for male adolescents this effect was found in ‘paternal’ households. While potential drawbacks in terms of generalisability and statistical power are recognised, other researchers are recommended, where appropriate, to consider using context-specific household definitions.

6.2. Introduction

Adolescence is a key period of biological and social development and events and experiences during this time can have long-ranging effects into adulthood. Living arrangements in adolescence in sub-Saharan Africa has been shown to be associated with schooling outcomes (Adjiwanou et al., 2021; Ijadunola et al., 2017), timings of transitions to adulthood, including sexual debut (Ngom et al., 2003; Pilgrim et al., 2014; Shoko et al., 2018; Tenkorang and Adjei, 2014), pregnancy (Shuvai Chikovore and Sooryamoorthy, 2022), marriage (Chae et al., 2016) and transition to the labour market (De Wet, 2012; Yamauchi et al., 2008), plus health-related outcomes such as engagement in risky sexual behaviours (Somefun, 2020) and emotional health (Wild, 2018). Living arrangements in adolescence may be more likely to be complex and changing as they tend to have high rates of migration and residential mobility (Beegle and Poulin, 2013; Grieger et al., 2013). Rural sub-Saharan Africa is traditionally very family orientated, but recent rapid social changes related to urbanisation, increased access to school and work opportunities, and use of smartphones and the internet may have an effect of how adolescence are living, and how they are affected by it.

The references cited above all use different ways to measure and define living arrangements, but most use the term ‘household’ in some way to define their exposure. This term is a commonly used unit of measurement across a wide range of sociological, demographic and epidemiological studies. Assigning individuals to households for analytical purposes is useful for avoiding double counting, generating sampling frames, and assessing non-individual interventions and exposures (Randall et al., 2015). However, there is no universal, standardised definition for household and its meaning varies across different settings and cultures (e.g. Van de Walle 2006, and references therein). To enable comparisons over time and contexts, and for ease, data are grouped in ways which may not reflect an individual’s experience, such as requiring an individual to associate with one household only (Randall et al., 2011), or conflating household membership with residence in

a particular homestead (Hosegood et al., 2005). There have also been suggestions that classifying African data into closed ‘households’ is inappropriate (Hertrich et al., 2020; Randall and Coast, 2015), though not all researchers agree fully with this sentiment (Rabe, 2008).

The literature around how to describe and define household groupings in a meaningful way has a long history across several disciplines including anthropology and demography (see Yotebieng and Forcone 2018 for a comprehensive summary). It has long been recognised that household composition should be studied in culturally sensitive ways (Yanagisako, 1979), and a range of methods have been used to improve household definitions in specific contexts, for example, some studies do use definitions based on ethnographic information or qualitative research (e.g. Breton 2019; Madhavan et al. 2017). Other studies, however, either bypass the issue by looking at the presence/absence of specific relatives (Ntshebe et al., 2019; Pilgrim et al., 2014; Zimmer, 2009) or use quite simplistic definitions such as whether a child lives with parents vs. no parents (Perkins, 2019) or whether a household includes a nuclear vs. extended family (Akinyemi et al., 2016). The nuclear family, which includes only parents and children, is often inappropriately held up as the ‘ideal’ even if it is not meaningful in other contexts, which not only makes data interpretation difficult, but also perpetuates western-centrism which has damaging consequences beyond research conclusions (McEwen, 2017; Sear, 2021). Simplistic definitions also fail to capture nuance and diversity. For example, the term ‘extended family’, which implies a core family plus extensions, is used widely but may lack validity in Africa (Siqwana-Ndulo, 1998) and other settings, where extended family could cover a wide range of living arrangements which may be very different from each other. There have also been some attempts to adapt simplistic definitions to be more culturally appropriate, though still tending to be simple, for example elementary (including nuclear, single parent and polygynous), three-generational and laterally extended households (Gage et al., 1997).

Detailed ethnographic study is not available or possible in many areas so researchers wanting to examine the effects of household composition must use these potentially flawed pre-existing ‘standard’ definitions or create their own, which can be complex and affected by investigator bias. In this study it is investigated whether, compared to standard definitions, a data-driven approach applied to secondary data improves understanding of the extent and frequency of variation in household composition among adolescents. 12 years of longitudinal data from a health and demographic surveillance site (HDSS) in northern Malawi are used, drawing on detailed information on relationships between individuals, to investigate different ways to define household membership (which household people are assigned to) and

household composition (the relationships between household members), including presenting a detailed assessment of the performance of latent class analysis (LCA).

In brief, these questions are addressed:

1. Is latent class analysis useful for creating household composition definitions?
2. Are the conclusions of sociological analyses different if household composition definitions:
    a. are data driven rather than using traditional, western-centric categories?
    b. include family members living very close by?

6.3. Methods

6.3.1. Context

The Karonga Health and Demographic Surveillance Site (HDSS) was established in 2002 in the southern part of the Karonga district in northern Malawi (Crampin et al., 2012). It covers an area of 150km^2 and by 2016 had over 40,000 people under surveillance, with very high response rates. Births and deaths are captured monthly through a system of local 'key informants', while migrations are captured annually through visits to all households. The area is largely rural with one semi-urban trading town, several smaller market villages and one port on Lake Malawi. The majority of the population engage in subsistence farming or fishing. The main ethnic group are Tumbuka, who have followed patrilineal and patrilocal custom since the 19th century: women tend to move to their husband's village when they marry (Malawi Human Rights Commission, 2006). In the event of divorce or even paternal death, children considered to be old enough to be away from their mother may be required to live with their father's family (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.

6.3.2 Dataset

Continuous data are available for participants of the HDSS for all the time they are living there. A simplified dataset was created which included one data point per year (15 June each year) per person. The main analyses use data from 2016 (as the latest year with sufficient completeness), however snapshots from 2004, 2007, 2010 and 2013 were also used to assess the repeatability and consistency of LCA output. Adolescents aged 12-18 (inclusive) were eligible for inclusion but those who were already married or had a child were

excluded. Adolescents who were not linked to their parents' identifiers (see below) were also excluded as without this information it was not possible to determine the nature of their relationship with other household members.

Unless otherwise specified in the text, all data processes and analyses were carried out using Stata 16.1.

6.3.3. Household membership definitions

In the HDSS, household membership is defined by the participants with guidance from trained fieldworkers: all household members must usually live in the dwelling/compound together and recognise 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 be living in each wife's household; all other individuals may only belong to one household. GPS coordinates are recorded for each household. Households are identified by a unique number which is linked to the household members' identification numbers. For this analysis, these household groupings are referred to as the 'immediate' household.

From previous field worker reports and data interrogation, it was known that two or more households sometimes reside in very close proximity (such as in the same compound), with shared facilities/resource and loose economic or social alliances. These grouped households were identified to generate an 'expanded' household definition. Grouped households are not formally identified during surveillance so a data-driven approach was used. As density of households varies across the area, to avoid erroneously linking unconnected households, and missing households that should have been connected, the decision to group households was based on local household density within 50m and how far apart the households are. In a densely populated area only very close households would be considered linked (within 5 metres) while in very rural areas where dwellings are more spaced, more distant households may be linked (up to within 20 metres): only households with at least one familial link (through blood or marriage) between household members were grouped.

6.3.4. Identification of relationships between household members

When a new household member 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 (even if they are not currently HDSS participants). 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. This information was used to identify all family links (by blood and by marriage) between all HDSS participants. Variables were generated to indicate whether a person was living with each family member (either in their ‘immediate’ household or using the ‘expanded’ definition described above).

6.3.5. Development of household composition variables

The processes used to create the different household composition (how the members are related or the ‘type’ of household) variables is described below. The household composition is described from point of view of the adolescent (i.e. only relationships between them and the other members are considered). In all cases, the unit of analysis is the individual adolescent, and each is treated as an independent data point even if they are living in the same household as another adolescent.

Traditional, western-centric household composition variables A ‘traditional’ variable was created to compare the data-driven definitions to, based on definitions commonly used in the literature: nuclear (both parents present and only under 18 siblings), extended (both parents present plus others), blended (one parent and one step-parent present), single parent (only one parent and no step-parent present) and no parents (no biological parent present). To emulate a common situation found in the literature when only Nuclear/Extended family is used, the ‘traditional’ variable was also further simplified to nuclear vs. non-nuclear, where non-nuclear includes extended, blended, single parent and no parents (this nuclear/extended dichotomous variable is referred to as the ‘basic’ variable).

Data-driven variables Creating the data-driven variables involves independent choices which are described in some detail to enable readers to assess the value of the process for themselves. The descriptions of the groupings found also provide useful and interesting information about the complexities of household composition in this setting.

Description of Latent Class Analysis

Latent class analysis (LCA) is a statistical technique which groups observations in otherwise unobserved classes (here, the household composition), based on a set of categorical variables (here, the presence or absence of types of relatives). For each observation, the probability of membership to each latent class is calculated before it is assigned to the group for which it has the highest value (maximum probability assignment rule).

The variables that were used in the LCA were chosen initially according to how common they were in the 2016 dataset, and their relevance in the local context and for the planned analyses. In general, if a relative type was living with at least 5% of adolescents it was included, however some less common relatives were included as they were felt to be important e.g., step-father (<2% of adolescents in 2016). The household composition list was finalised following preliminary attempts at the LCA: it was kept as simple as possible by combining or removing variables which did not seem to have an effect on the classes found (i.e., the groupings were similar with or without the relative type). In the final analyses the following relatives were included: mother, father, maternal grandparent, paternal grandparent, father's wife (not mother), mother's husband (not father), half-sibling under-18 (maternal), half-sibling under-18 (paternal), full sibling under-18, sister over-18 (half or full), brother over-18 (half or full); along with the following relative categories: maternal family under 18, maternal family over 18, paternal family under 18, paternal family over 18 (these groups include all maternal or paternal family not listed above including aunts, uncles, cousins etc.), brother's family and sister's family (these groups include the half or full sibling's children or grandchildren).

For the 'immediate' household analyses, all input variables were binary (none vs. at least one of the relatives) as few differences were found when categorical (none, few [less than the median], many [above the median]) variables were used. For the 'expanded' household analyses, the variables were coded as the relative not present, present in 'immediate' household only, present in both 'immediate' and 'expanded' household, and only present in the 'expanded' household.

To assess the repeatability and consistency of the LCA results, the same analyses were carried out separately using data from 2004 (n=3175), 2007 (n=3683), 2010 (n=3821), 2013 (n=4381) and 2016 (n=5364). The analysis was run separately using the 'immediate' household and the 'expanded' household membership definitions for 3-15 (inclusive) latent classes using the poLCA (Linzer and Lewis, 2011) package in R. The solution selection was guided by the Bayesian information criterion (BIC), as this has been shown to be the most reliable measure (Nylund et al., 2007). However, when the BIC-optimal solution resulted in groups smaller than 100 individuals, another solution with a similar BIC, entropy and average probability of assigned group membership was chosen to try to ensure stability of the models (Nylund-Gibson and Choi, 2018). For example, the 8-class solution for the 2016 'immediate' household had the optimal BIC, however the 7-class solution was selected as it had a similar BIC, similar entropy (81.3% vs. 79.4%) and a similar average likelihood of group membership (92.3% vs. 90.5%) but had 119 in the smallest group rather than 75.

Results of Latent Class Analysis

The latent classes for the 'immediate' household composition are described below. Although the LCAs on the 5 year-specific datasets were run independently so the classes are not comparable, many were similar enough to be described with the same name. Any variations in the classes are described below and unless otherwise stated, each class is present in each of the 5 year-specific analyses. The relatives not mentioned in each class have a very low/no (<0.1) chance of being present. The probability of membership of each class is calculated for each observation as part of the LCA process: for many of the classes the average probability across all the members assigned to that class was high (over 90%) implying that one can be sufficiently confident that people are assigned to an appropriate class. However, some classes had a slightly lower average probability (indicated in the text, full tabulation for the 2016 analysis is found in table 6.1).

'Immediate' household latent classes:

1. ***Parents & siblings***: this class has very high probability of presence of the adolescent's mother, father and full sibling(s) aged under 18 (however none are 1.0, and few of the other relatives are at 0, though are all low). This class was fairly discriminatory: in 2016 over 75% of members had over 90% probability of being in the class, and only 1.3% had less than 60% probability. In 2016, 55.9% of adolescents lived in these households (the largest class).
2. ***Parents & sister's family***: this class has a very high probability of the adolescent's mother, high probability of their father and full sibling(s) aged under 18 and high/very high chance of their sister(s) and family, there is also some chance of their brother(s). The probability of the adolescent's sister, their sister's family and their brother varies in each year (i.e. probability of 'sister over 18' varies from 0.58 to 0.74). In 2016, 13.9% of adolescents lived in these households.
3. ***Brother's family***: this class has very high likelihood of the presence of the adolescent's brother(s) aged over 18 and good chance of their brother's family and their mother. It did not appear in the 2010 and 2013 analyses and the 2016 version has a much lower chance of 'brother's family' than the 2004 and 2007 versions (0.84 in 2004 vs. 0.53 in 2016). This class had the lowest discriminatory power, only 37% of class members had over 90% chance of being in the class and 13.4% had a less than 60% chance (most of these had next likeliest category of 'parents and sister's family'), however the average likelihood was still quite high at almost 80%. This was the smallest group in 2016 with 2.2% of adolescents.
4. ***Mother & siblings***: this class has a very high probability of the adolescent's mother and low/no chance of their father, chance of maternal siblings is usually high or very

high, but chance of ‘mother's husband (not father)’ is low. In 2016, 100% of class members lived with their mother, however in other years this was as low as 81%. This was the second smallest group in 2016 with 5.5% of adolescents.

5. ***Father & step-mother***: this class has a very high or high probability of the adolescent’s father, their father's wife and paternal siblings but probability of their full siblings and older brother or sister is lower. The probabilities of these key relatives are similar across the datasets, however in some years there is a low chance of presence of their biological mother as well, which would mean that some of this class were actually polygynous households. This class had a very high average probability of membership (98.6% in 2016). In 2016, 8% of adolescents lived in this type of household.
6. ***Maternal***: this class has good/high probability of the adolescent’s maternal grandparents, their maternal family aged under and over 18, and low chance of their mother, full siblings and maternal siblings. In 2016, 8.1% of adolescents lived in this type of household.
7. ***Paternal***: this class has good/high probability of the adolescent’s paternal grandparents, their paternal family aged under and over 18, and low chance of their full siblings. In some years there was a low chance of their paternal siblings and father. In 2016, 6.5% of adolescents lived in this type of household.

The analyses including the ‘expanded’ household produced some classes that were similar to the ‘immediate’ household classes (‘Parents & siblings’, ‘Parents, siblings and sister’s family’, ‘Mother & siblings’, ‘Father and step-mother’, ‘Maternal’ and ‘Paternal’). There were also 4 additional classes, which are described below. Note that the -> signifies ‘expanding to’, i.e. ‘Parents & siblings-> paternal’ means parents and siblings family in the ‘immediate’ household plus paternal family in the ‘expanded’ household. The average probability of membership tended to be higher with this analysis with only one class having less than 90% and 6 out of 10 classes having over 95% (table 6.1).

’Expanded’ household latent classes:

1. ***Parents->brother's family***: this class had a very high probability of the adolescent’s mother, high probability of their father and full sibling(s) aged under 18 in the ‘immediate’ household, and very high probability of their brother's family in the ‘expanded’ household. There is usually a chance of their brother(s) aged over 18 and their brother's family in the ‘immediate’ household as well, and their sister(s) and sister's family in the ‘immediate’ household.

2. ***Parents & siblings->paternal***: this class had a high/very high probability of the adolescent's parents and full siblings under 18 in the 'immediate' household, and high/very high probability of their paternal family aged under 18 and over 18 in 'expanded' household. There is also low chance of their older siblings in the 'immediate' household and of their paternal grandparents in the 'expanded' household.
3. ***Mother & siblings->maternal***: In this class the probability of the adolescent's mother is good/high and there is a very high probability of maternal relatives in the 'expanded' household (low chance of maternal relatives in the 'immediate' household). This class only appears in 2013 and 2016.
4. ***Polygynous***: this class has a very high probability of the adolescent's mother and full siblings under 18 in 'immediate' household and very high chance of their paternal sibling(s) and father's wife in 'expanded' household, high chance of their father in both the 'immediate' and 'expanded' household, low chance of their older siblings in the 'immediate' and 'expanded' household and of paternal family in the 'expanded' household.

The classes found through both the 'immediate' household and 'expanded' household analyses are distinct from each other and contextually appropriate. The LCA statistics and the repeatability (independent LCA conducted on datasets from the 5 separate years produce similar looking classes and some similar classes found with both the 'immediate' and 'expanded' household analyses) give confidence in the class assignments and the technique.

Investigator input was required for choosing and coding the input variables, choosing the number of classes and interpreting and labelling them. Low probabilities of relatives in some classes complicated the labelling process, i.e. the small probability of also finding a mother in the 'father and step-mother' class would mean that this class might include some polygynous households where the 'step-mother' is a co-wife. Equally, uncertainty existed even in classes with high probabilities, i.e., in 2016 the 'parents & siblings' class has 0.98 chance of 'mother' and 0.85 chance of 'father' so while having a high chance of both parents is a defining feature of the class, clearly not all members fulfil this definition.

Due to these complexities the investigators chose to create LCA-guided household composition variables with similar categories to the latent classes, however as few of the latent classes had 100% chance of presence of the key relatives it was not possible to replicate them entirely. The logical rules used to create the categories are found in table 6.2.

Table 6.1: Distribution and likelihood of group membership for the latent classes found

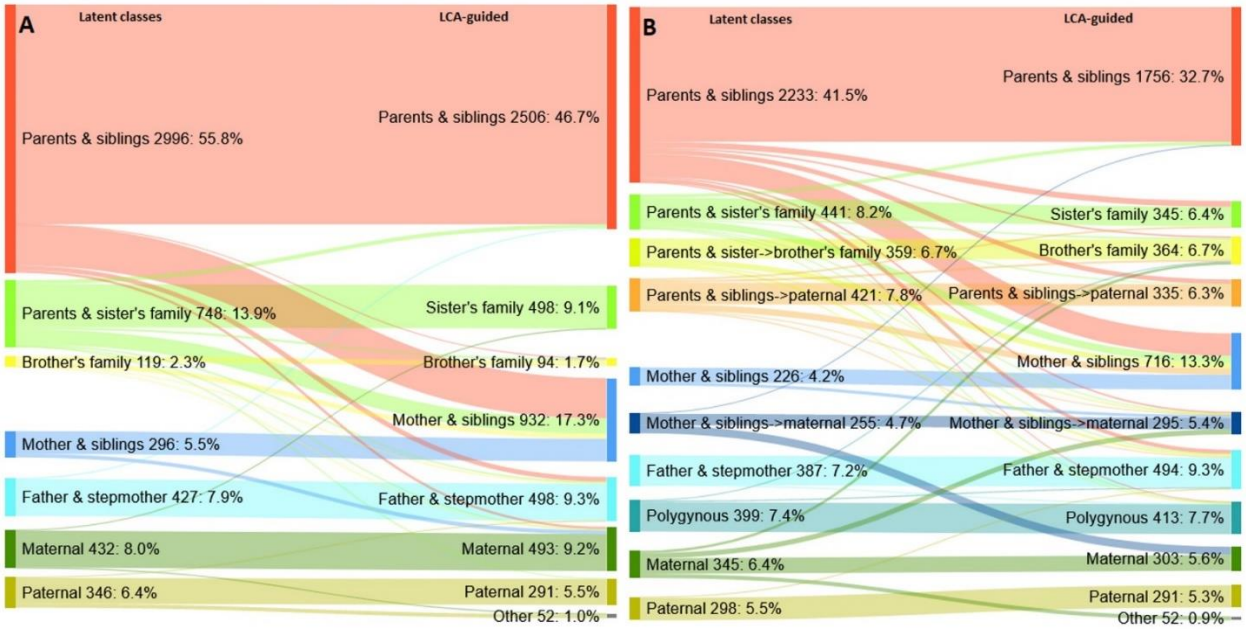
			Likelihood of class membership							Next likeliest class if p<60%	
	N	%	Mean	SD	Minimum	p >90%		p <60%		Class	N
						N	%	N	%		
Not included in LCA	1824	25.4%									
Included in LCA	5364	74.6%									
Immediate household data only											
Parents & siblings	2996	55.9%	92.9%	10.0%	27.9%	2271	75.8%	39	1.3%	Parents & sister's family	30
Parents and sister's family	748	13.9%	88.2%	13.3%	40.9%	467	62.4%	11	1.5%	Parents & siblings	9
Brother's family	119	2.2%	79.5%	19.3%	36.4%	44	37.0%	16	13.4%	Parents & sister's family	10
Mother and siblings	296	5.5%	87.1%	19.0%	37.5%	206	69.6%	33	11.1%	Parents & siblings	29
Father & step-mother	427	8.0%	98.6%	6.8%	39.3%	414	97.0%	5	1.2%	Parents & siblings	3
Maternal	432	8.1%	95.7%	12.3%	33.9%	374	86.6%	16	3.7%	Paternal	8
Paternal	346	6.5%	92.7%	17.5%	43.7%	295	85.3%	45	13.0%	Maternal	41
Expanded household data added											
Parents and siblings	2233	41.6%	93.5%	11.7%	40.6%	1852	82.9%	89	4.0%	Parents and sister's family	57
Parents & sister's family	441	8.2%	87.0%	12.8%	42.0%	200	45.4%	23	5.2%	Parents & siblings	15
Parents->brother's family	359	6.7%	99.3%	3.9%	50.2%	350	97.5%	1	0.3%	Maternal	1
Parents->paternal	421	7.8%	99.2%	3.9%	45.7%	412	97.9%	1	0.2%	Parents & siblings	1
Mother and siblings	226	4.2%	90.4%	16.1%	40.0%	169	74.8%	20	8.8%	Parents & siblings	13
Mother and siblings->maternal	255	4.8%	97.1%	9.0%	47.5%	234	91.8%	6	2.4%	Parents & siblings	3
Father & step-mother	387	7.2%	97.6%	8.5%	31.0%	360	93.0%	7	1.8%	Parents & siblings	3
Polygynous	399	7.4%	99.7%	2.8%	51.0%	396	99.2%	1	0.3%	Parents & siblings	1
Maternal	345	6.4%	93.3%	10.2%	45.7%	252	73.0%	5	1.4%	Parents & siblings	3
Paternal	298	5.6%	99.0%	5.3%	39.5%	292	98.0%	2	0.7%	Parents & siblings	2

Table 6.2: Logical rules used to create LCA-guided categories

'Immediate' household categories	'Expanded' household categories
Parents & siblings: Both parents present and does not fit into any of the non-'*other'* categories	*Parents & siblings*: Both parents present in immediate household, brother and family and paternal family not present in expanded household and does not fit into any of the non-'*other'* categories
Sister's family: At least 1 over-18 sister or her family, sister+family larger than brother+family, and mother or father present or no maternal or paternal family present	*Sister's family*: At least 1 over-18 sister or her family, sister+family larger than brother+family, and mother or father present or no maternal or paternal family present, no brother and/or family in the expanded household
Brother's family: as above but with brother instead of sister	*Brother's family*: At least 1 over-18 brother or his family, brother+family larger than sister+family, and mother or father present or no maternal or paternal family present, or brother and/or family in the expanded household
	Parents & siblings->paternal: Both parents present in immediate household, paternal family present in expanded household and does not fit into any of the other non-other categories
Mother & siblings: mother present, no father, father's other wife nor maternal family	*Mother & siblings*: mother present, no father, step-mother or maternal family in expanded household
	Mother & siblings->maternal: mother present, no father, step-mother in immediate or expanded household, maternal family in expanded household
Father & stepmother: mother not present, father or father's other wife present	*Father & stepmother*: mother not present, father or step-mother present
	Polygynous: mother and another father's wife present in immediate or expanded household
Maternal: father and father's other wife not present, at least 1 maternal relative present and maternal relatives larger than paternal	*Maternal*: No mother, father nor other father's wife, at least 1 member of maternal family in immediate or expanded household and maternal family larger than paternal family
Paternal: mother and father's other wife not present, at least 1 paternal relative present and paternal relatives larger than maternal	*Paternal*: as above but with paternal rather than maternal
Other: Does not fit into any of the above categories	*Other*: Does not fit into any of the above categories

The correspondence between the latent classes and the LCA-guided categories using the 2016 data is shown Sankey diagrams in figure 6.1. These diagrams show how data flow between categories, with the width of the band connecting the categories indicating the number of observations (created using http://sankeymatic.com/). Although there is good correspondence between many categories like 'father and step-mother', 'maternal', 'paternal' and 'polygynous', the 'mother & siblings' LCA-guided category is larger, taking a proportion of 'parents & siblings' and 'parents and sister's family'.

Figure 6.1: Sankey diagrams showing correspondence between latent classes (left) and LCA-guided categories (right) for A. 'immediate' household only and B. 'expanded' household added (n=5364)



NB. Definitions for the LCA-guided categories are shown in table 6.2

6.3.6. Description and comparison of household composition definitions

Sankey diagrams and simple tabulations were used to compare the 'traditional' definitions with the LCA-guided ones and to examine the changes in the categories the adolescents fell into when using the 'immediate' and 'expanded' household LCA-guided definitions. Only data from 2016 were included in these analyses.

6.3.7. Example statistical analyses using the different household composition definitions

To assess the utility of the different household composition definitions in sociological analyses, the associations between the variables and three simplified outcomes relating to

adolescence were examined with logistic regression models using the 2016 data. The outcomes were:

1. Schooling: In Malawi access to secondary education is still challenging (only 38.4% of pupils transitioned from primary to secondary school in 2017/18, and only 24% completed secondary school (National Statistical Office, 2019)). All HDSS participants are asked annually about their school enrolment status, including current grade if attending school. For this analysis, a simple outcome of whether the adolescent was currently behind or dropped out of school was used. In this dataset the median age for grade was 2 years older than expected, thus for this outcome 2 or less years older than optimal is classed as appropriate and 3 or more years older is classed as 'behind'.
2. Child-bearing: Marrying and/or having a child before the age of 18 has been relatively common in this area, although 'child marriage' was outlawed in 2017 (Daniel, 2017). Child-bearing during adolescence is associated with school drop-out (Glynn et al., 2018) and later life health issues (Cho et al., 2012). This analysis uses a simple outcome of living with biological child in the following year (2017); it was restricted to women who were still in the area in 2017.
3. Weight: Being underweight during adolescence can have ill effects into adulthood and can prevent people from reaching their full potential. HDSS participants have regularly been weighed and measured as part of various surveys. For this analysis the nearest measurements taken within 1 year of the data point was taken and the outcome was whether the adolescent was classed as underweight (weight for height at least one standard deviation below WHO standards).

Models were run separately for male and female adolescents and each household composition variable was added separately to a baseline model including socio-economic and demographic factors. These *a priori* variables were chosen based on known associations and data availability:

- whether mother and father had secondary education
- distance to the nearest school (<0.5km, 0.5-1km and over 1km)
- 2 proxies for current household economic status:
    - source of water (household tap, shared tap, bore hole or well/other)
    - a score of assets owned by the household. For the latter, the total number of low value possessions was added up (sofa, mattress, bed, bicycle, table, paraffin lamp, chair and radio) and transformed into a quartile score variable

relative to the other data, the total number of high value (ox-cart, fridge, car, motorbike or television) assets was then added to this score.

- age of adolescent
- number of adults and number of children in the household

Standard errors were calculated allowing for intragroup correlation within households, as some adolescents in the dataset may be living together. It should be noted that although there was an attempt to control for confounding factors, the aim of these analyses was to compare the impact of using different household composition variables rather than fully assess the causal relationship between household composition and the outcomes.

6.4. Results

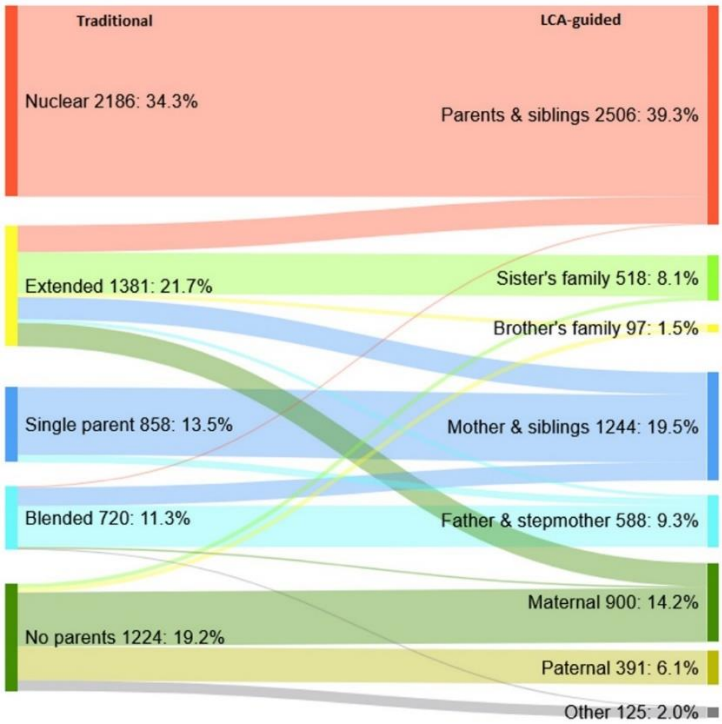
There were 9343 households in the HDSS area in 2016, 4444 (47.5%) included at least one adolescent aged 12-18 inclusive. There were 7188 adolescents in total, 411 were already married/had children (22 male, 389 female) so were excluded. A further 408 were not linked to either parent identifier and were also excluded, leaving 3507 male, and 2862 female participants. Just under 40% of the included adolescents' households were linked to at least one other household for the 'expanded' household composition variable: 23.8% to one household, 8.3% to two other households and 4.8% to three or more, this was similar for both females and males and at different ages.

6.4.1. Comparing 'traditional' with LCA-guided household composition definitions ('immediate' household only)

Using the 'traditional' household composition definitions and the 'immediate' household membership definition, the largest group of adolescents lived in a 'nuclear' household, however the percentage was relatively low at 34.3%, the other categories appeared in order of size: 'extended' (21.7%), 'no parent' (19.2%), 'single parent' (13.5%) and 'blended' (11.3%). The largest category using the LCA-guided definition was 'parents & siblings' (39.3%), followed by 'mother & siblings' (19.5%), 'maternal' (14.2%), 'father & stepmother' (9.3%), 'sister's family' (8.1%), 'paternal' (6.1%), 'other' (2%) and 'brother's family' (1.5%). The correspondence between these two definitions is shown in a Sankey diagram in figure 6.2. There is not a one-to-one relationship between any of the categories, showing that the LCA-guided variable is not simply a more complex version of the traditional variable. Almost the whole traditional 'nuclear' category are classed as the LCA-guided 'parents & siblings' category, however as the 'parents & siblings' group allows for other relatives being present,

part of the traditional ‘extended’ category are also included. The traditional ‘extended’ category includes most members of the ‘sister’s family’ and ‘brother’s family’ groups, and part of the ‘maternal’ group. The traditional ‘blended’ group includes most of the ‘father and step-mother’ category, and part of ‘mother & siblings’. The latter also includes most of the ‘single parent’ category’. The ‘paternal’ and ‘other’ categories are almost totally found in the traditional ‘no parents’ category, along with the majority of the ‘maternal’ group.

Figure 6.2: Sankey diagram showing the correspondence between the 5-level ‘traditional’ (left) and LCA-guided (right) household composition definitions (‘immediate’ household only, n=6369)



NB. Definitions for ‘traditional’ categories found in the methods and for the LCA-guided categories in table 6.2

6.4.2. Comparing ‘immediate’ household and ‘expanded’ household composition definitions

The correspondence between the LCA-guided ‘immediate’ household definitions and the ‘expanded’ household definitions is shown in table 6.3. In total, 1462 (23%) adolescents were classified differently when using the ‘expanded’ household definitions compared to when using the ‘immediate’ household. Some categories were almost the same when using the 2 definitions (‘father and step-mother’, ‘paternal’) while ‘parents & siblings’, ‘sister’s family’, ‘mother & siblings’ and ‘maternal’ had more differences. Those in ‘parents & siblings’ households using the ‘immediate’ definition commonly were classified as ‘brother’s family’,

‘parents & siblings ->paternal’ or ‘polygynous’ using the ‘expanded’ one; those in the ‘sister’s family’ ‘immediate’ category were most commonly categorised as ‘brother’s family’ using the ‘expanded’ definition; those in the ‘immediate’ ‘mother & siblings’ often were categorised as ‘mother & siblings->maternal’ or ‘polygynous’ in the ‘expanded’ definition; and ‘immediate’ ‘maternal’ were most commonly regrouped as ‘mother & siblings->maternal’ in the ‘expanded’ version.

The value of including very close relatives is demonstrated here, providing more evidence for the complexity of living arrangements as even fewer adolescents can be categorised as living in a ‘nuclear’-type family (‘parents & siblings’), and also aids in distinguishing between those living in households which may be vulnerable (i.e. ‘mother & siblings’) and those in that category who may be less vulnerable due to proximity to other support networks (i.e. ‘mother & siblings->maternal’).

Table 6.3: Correspondence between ‘immediate’ and ‘expanded’ household categories using the LCA-guided household composition definition

‘Expanded’ household added	‘Immediate’ household only								Total
	Par &sib	Sis fam	Bro fam	Moth &sib	Fath &st	Mat	Pat	Oth	
Parents & siblings	**1,756**	*0*	*0*	*0*	*0*	*0*	*0*	*0*	1,756
Sister's family	*0*	**351**	*0*	1	*0*	*0*	*0*	*0*	352
Brother's family	170	103	**94**	*0*	*0*	*0*	*0*	1	368
Parents & siblings-> paternal	335	*0*	*0*	*0*	*0*	*0*	*0*	*0*	335
Mother & siblings	*0*	*0*	*0*	**989**	*0*	*0*	*0*	*0*	989
Mother & siblings-> maternal	*0*	33	*0*	132	*0*	274	*0*	*0*	439
Father & stepmother	*0*	*0*	*0*	*0*	**584**	*0*	*0*	*0*	584
Polygynous	245	31	3	122	4	8	*0*	*0*	413
Maternal	*0*	*0*	*0*	*0*	*0*	**618**	*0*	*0*	618
Paternal	*0*	*0*	*0*	*0*	*0*	*0*	**391**	*0*	391
Other	*0*	*0*	*0*	*0*	*0*	*0*	*0*	**124**	124
Changed	750	167	3	255	4	282	0	1	1,462
%	29.9%	32.2 %	3.1 %	20.5 %	0.7 %	31.3 %	0.0 %	0.8 %	23.0 %

NB. Bolded numbers show no change in category from immediate to expanded household; definitions for each category can be found in table 6.2

6.4.3. Example statistical analyses using the household composition definitions

All results in this section come from logistic regression models adjusted for socio-demographic factors and household size. The baseline group was ‘nuclear’ for the basic and ‘traditional’ variables, and ‘parents & siblings’ for the LCA-guided variables. As the aim of this section is to assess whether analytical conclusions would be different if different household variables are used the actual odds ratios found are not quoted, however all are available in tables 6.4-6.6.

Education outcome

The education outcome was missing for 275 adolescents with 2119 (34.8%) of the remaining 6093 adolescents with sufficient household data classed as behind or dropped out of school. A further 600 did not have sufficient data in all of the variables of the regression model. There was no evidence of an association between household composition and the education outcome when using the basic definition (adolescents categorised as living in either ‘nuclear’ or ‘non-nuclear’ households), however when using the ‘traditional’ 5-level household composition definitions, for both female and male adolescents the ‘blended’ and ‘no parents’ categories were associated with increased likelihood of being behind or dropped out of school. With the LCA-guided ‘immediate’ household definition, both female and male adolescents living in the ‘father and step-mother’ category had higher odds of experiencing the outcome, which agrees with the ‘traditional’ variable findings, however only female adolescents living in ‘maternal’ households had higher odds of the outcome while male adolescents living in ‘paternal’ households had higher odds. When the ‘expanded’ household definition was used, there was some evidence that female adolescents living in ‘Brother’s family’ households had higher odds of the outcome (table 6.4).

Child-bearing outcome

This outcome was missing for 168 female adolescents: 116 (4.3%) of the remaining 2693 female adolescents with sufficient household data were living with their own child in 2017. A further 292 did not have sufficient data in all of the variables of the regression model. There was evidence that adolescents in non-nuclear households using the binary variable had a lower odds of the outcome and using the ‘traditional’ variable showed that the odds were lower for ‘blended’ and ‘single parent’ households. Using the LCA-guided variable showed the odds were only lower for ‘father and step-mother’ households and using the ‘expanded’ household variable showed evidence of increased odds for those in ‘parents & siblings-> paternal’ households, compared to those living with parents and siblings but not close to paternal family (table 6.5).

Table 6.4: Logistic regression of the association between the different household composition variables and odds of poor educational outcome (being behind or dropped out of formal education)

	Female (n=2605)				Male (n=3157)			
	aOR#	p	95% CI		aOR#	p	95% CI	
Nuclear vs. non-nuclear ('immediate' household only)								
Nuclear	reference				reference			
Non-nuclear	1.2	0.198	0.9	1.5	1.2	0.150	1.0	1.4
'Traditional' ('immediate' household only)								
Nuclear	reference				reference			
Extended	1.1	0.643	0.8	1.4	1.0	0.725	0.7	1.2
Blended	**1.7**	**0.003**	**1.2**	**2.4**	**1.3**	**0.046**	**1.0**	**1.8**
Single parent	0.8	0.160	0.5	1.1	1.1	0.475	0.8	1.5
No parents	**1.6**	**0.006**	**1.1**	**2.3**	**1.5**	**0.005**	**1.1**	**1.9**
LCA-guided ('immediate' household only)								
Parents & siblings	reference				reference			
Sister's family	1.1	0.747	0.7	1.5	0.9	0.450	0.6	1.2
Brother's family	0.8	0.661	0.3	2.1	1.5	0.184	0.8	2.9
Mother & siblings	0.7	0.069	0.5	1.0	1.0	0.950	0.8	1.3
Father & stepmother	**1.9**	**0.001**	**1.3**	**2.7**	**1.4**	**0.019**	**1.1**	**1.9**
Maternal	**1.7**	**0.003**	**1.2**	**2.5**	1.1	0.632	0.8	1.5
Paternal	1.6	0.086	0.9	2.8	**1.5**	**0.025**	**1.1**	**2.2**
Other	0.6	0.146	0.3	1.2	1.1	0.809	0.6	1.9
LCA-guided ('expanded' household added)								
Parents & siblings	reference				reference			
Sister's family	0.8	0.406	0.5	1.3	0.8	0.355	0.6	1.2
Brother's family	***1.6***	***0.055***	***1.0***	***2.4***	1.4	0.091	0.9	2.0
Par & sibl-> paternal	1.4	0.241	0.8	2.2	1.4	0.118	0.9	2.0
Mother & siblings	1.0	0.811	0.7	1.4	1.1	0.490	0.8	1.5
Moth & sibl-> maternal	1.3	0.352	0.8	2.1	1.0	0.924	0.6	1.6
Father & stepmother	**2.1**	**<0.001**	**1.4**	**3.2**	**1.6**	**0.003**	**1.2**	**2.1**
Polygynous	1.1	0.593	0.7	1.7	1.4	0.079	1.0	2.0
Maternal	**2.3**	**<0.001**	**1.5**	**3.5**	1.4	0.105	0.9	2.0
Paternal	**1.9**	**0.032**	**1.1**	**3.3**	**1.7**	**0.006**	**1.2**	**2.5**
Other	0.9	0.493	0.5	1.5	1.2	0.461	0.7	2.2

NB. Results with p-value<0.05 are marked in bold and those with p<0.075 are in italic bold; definitions for all categories are shown in the methods section or table 6.2; #adjusted OR: each model included parental education level, distance to nearest school, household economic status, age of adolescent and number of adults and children in the household.

Table 6.5: Logistic regression of the association between the different household composition variables and odds of having a child in next year

	Female (n=2530)			
	aOR#	p	95% CI	
Nuclear vs. non-nuclear				
Nuclear	reference			
Non-nuclear	***0.7***	***0.068***	***0.4***	***1.0***
'Traditional' (immediate household only)				
Nuclear	reference			
Extended	0.9	0.711	0.5	1.5
Blended	**0.4**	**0.029**	**0.1**	**0.9**
Single parent	**0.4**	**0.021**	**0.2**	**0.9**
No parents	0.7	0.345	0.3	1.5
LCA-guided (immediate household only) [n=2505]				
Parents & siblings	reference			
Sister's family	0.9	0.845	0.5	1.8
Brother's family	-	-	-	-
Mother & siblings	0.6	0.084	0.3	1.1
Father & stepmother	**0.3**	**0.045**	**0.1**	**1.0**
Maternal	0.7	0.322	0.3	1.5
Paternal	0.6	0.322	0.2	1.8
Other	0.7	0.662	0.1	4.0
LCA-guided ('expanded' household added)				
Parents & siblings	reference			
Sister's family	1.0	0.954	0.4	2.3
Brother's family	0.5	0.163	0.3	1.3
Par & sibl-> paternal	**2.3**	**0.026**	**1.1**	**4.8**
Mother & siblings	0.6	0.252	0.3	1.4
Moth & sibl-> maternal	0.4	0.123	0.1	1.3
Father & stepmother	0.4	0.082	0.1	1.1
Polygynous	1.0	0.954	0.5	2.3
Maternal	0.9	0.808	0.3	2.5
Paternal	0.6	0.435	0.2	2.0
Other	0.8	0.779	0.1	5.4

NB. Results with p-value<0.05 are marked in bold and those with p<0.075 are in italic bold; definitions for all categories are shown in the methods section or table 6.2; #adjusted OR: each model included parental education level, distance to nearest school, household economic status, age of adolescent and number of adults and children in the household.

Table 6.6: Logistic regression of the association between the different household composition variables and odds of being underweight (weight for height 1 standard deviation or more from WHO standards)

	Female (n=1529)				Male (n=1866)			
	aOR#	p	95% CI		aOR#	p	95% CI	
Nuclear vs. non-nuclear (immediate household only)								
Nuclear	reference				reference			
Non-nuclear	1.0	0.833	0.8	1.3	1.1	0.451	0.9	1.3
'Traditional' (immediate household only)								
Nuclear	reference				baseline			
Extended	1.2	0.381	0.8	1.6	1.1	0.714	0.8	1.4
Blended	1.2	0.515	0.8	1.8	1.0	0.981	0.7	1.4
Single parent	0.8	0.279	0.5	1.2	1.1	0.487	0.8	1.6
No parents	1.0	0.818	0.6	1.4	1.2	0.243	0.9	1.6
LCA-guided (immediate household only)								
Parents & siblings	reference				reference			
Sister's family	0.9	0.606	0.5	1.4	0.9	0.498	0.6	1.3
Brother's family	1.6	0.442	0.5	4.8	1.4	0.374	0.7	2.9
Mother & siblings	0.8	0.260	0.6	1.2	1.2	0.249	0.9	1.6
Father & stepmother	1.3	0.220	0.8	2.1	1.0	0.880	0.7	1.5
Maternal	**1.4**	***0.065***	**1.0**	**2.1**	**1.6**	**0.009**	**1.1**	**2.3**
Paternal	***0.5***	***0.054***	***0.2***	***1.0***	0.9	0.724	0.6	1.4
Other	0.7	0.290	0.3	1.4	0.6	0.134	0.3	1.2
LCA-guided ('expanded' household added)								
Parents & siblings	reference				reference			
Sister's family	0.8	0.379	0.4	1.4	1.0	0.854	0.6	1.5
Brother's family	1.0	0.965	0.6	1.8	1.2	0.425	0.8	1.8
Parents & siblings-> paternal	0.9	0.599	0.5	1.4	1.1	0.603	0.7	1.8
Mother & siblings	0.8	0.220	0.5	1.2	1.2	0.251	0.9	1.7
Mother & siblings-> maternal	**1.8**	**0.023**	**1.1**	**3.1**	1.2	0.382	0.8	1.9
Father & stepmother	1.4	0.194	0.8	2.3	1.1	0.780	0.7	1.5
Polygynous	1.0	0.893	0.7	1.6	1.0	0.864	0.7	1.4
Maternal	1.3	0.268	0.8	2.1	**1.9**	**0.003**	**1.2**	**2.9**
Paternal	***0.5***	***0.062***	***0.2***	***1.0***	1.0	0.844	0.6	1.4
Other	0.7	0.381	0.3	1.5	0.7	0.212	0.3	1.3

NB. Results with p-value<0.05 are marked in bold and those with p<0.075 are in italic bold; definitions for all categories are shown in the methods section or table 6.2; #adjusted OR: each model included parental education level, distance to nearest school, household economic status, age of adolescent and number of adults and children in the household.

Underweight outcome

This outcome was missing for 2738 adolescents with 1364 (37.6%) of the remaining 3630 adolescents with sufficient household data classed as underweight. A further 393 did not have sufficient data in all of the variables of the regression model. There are no observable associations when using the binary or the 'traditional' variable. However, using the LCA-guided variable for female adolescents there is evidence of increased odds in 'maternal' households but decreased odds in 'paternal' households. For males there is also increased odds in 'maternal' households. Using the LCA-guided expanded household definition shows that for female adolescents, the increased odds are specifically in the 'mother & siblings->maternal' households, and not in the 'maternal' households (table 6.6).

6.5. Discussion

This study used LCA, a data driven statistical method, to better understand the distribution and frequency of household composition compared to investigator-generated traditional methods. It attempted to build upon existing literature (which mostly uses additional primary data collection to improve household definitions (e.g. Madhavan et al. 2017)), by using a method applied to an existing dataset, from which recommendations are derived about how to improve data collection on household structures. LCA generated a wide range of household composition classes, of which many were distinct from those used in 'traditional' household composition definitions. These LCA household composition classes were appropriate to the context. The statistical methodology is robust, with high entropy values and most class members having high likelihood of membership. The LCA method was repeatable across datasets from the same area over calendar time and when using different household membership definitions. This data driven approach minimises some investigator biases, although the investigator can influence, to varying degrees, the selection of input variables and the total number and description of the final latent classes. To be useful for analysis, the classes must be labelled in a concise manner which may inadvertently mask the complexities and uncertainties within them (although the LCA process also served to demonstrate the complexities of household composition): it was found to be most useful to use the LCA results as a guide for creating household composition variables.

LCA is commonly used in analyses of sociological outcomes (e.g. parenting (Hwang and Jung, 2021), health care utilisation (Traino et al., 2021) etc.), yet there are few published examples of LCA being used to create household composition definitions in the literature (Huffman et al., 2019b; Lee et al., 2020; Liao, 2004). Comparable to the present analyses, two studies used age, sex and relationship status input variables(Huffman et al., 2019b; Lee

et al., 2020), however differences in research question and data availability meant that the final variables used and the classes found varied between studies; providing further evidence of the complexities involved in studying household composition. Another data-driven approach, sequence analysis, has been used to assign household composition categories to census data using relationship of each household member to the household head: the authors found this technique gave more informative categories than standard definitions used with census data, though the relationship data was not as detailed as in the present analysis (Bignami-Van Assche et al., 2021).

Using the 'expanded' household definition, where the household was not just restricted to those living under the same roof, revealed new and interesting categories and living arrangements that were not apparent from the 'immediate' household classification (for example polygynous households, and 'nuclear' families living very near paternal relatives). The limitations of analysing African 'households' as individual units in isolation have been described previously (Rabe, 2008). Furthermore, use of expanded household definitions (e.g. by conducting in-depth interviews following surveys to assess how well the survey definition performed (Kriel et al., 2014), including multiple definitions during the same survey (Beaman and Dillon, 2012), or using specially designed surveys to assess support networks (Madhavan et al., 2018, 2017a)) have been shown to provide more informative groupings. To the authors knowledge, this study is the first to expand household definitions beyond the residential unit using secondary data.

The findings from logistic regression analyses show that compared to the traditional definitions, the context appropriate definitions (informed by the LCA method), provide a more nuanced understanding of the relationship between household composition and other factors. Associations were observed between a. poor educational outcome for female adolescents in a 'maternal' household compared to those living with 'parents and siblings', while for male adolescents a similar effect was seen for 'paternal' households; b. child-bearing within next 12 months in female adolescents living with 'parents & siblings expanding to paternal' compared to those living with 'parents and siblings' and; c. being underweight in a maternal household compared to a 'parents and siblings' household for both male and female adolescents, with some evidence for an inverse association for underweight in females living in a paternal household (that was not observed for males) compared to a 'parents and siblings' household. All of these associations would not have been detectable using the 'traditional' household definitions. The logistic regression analyses were presented with the main aim of assessing the extent to which associations with a range of outcomes varied with different household composition variables, rather than trying to

understand the associations found. Although the logistic regression models included adjustment for several important socio-demographic variables, formal causal analysis was not attempted and speculation on the pathways or mechanisms of the associations found is not within the scope of this paper.

Limitations

This analysis benefitted from highly detailed longitudinal data, but there are a number of potential limitations to take into account.

The technique used to group households to generate the 'expanded' household definition may have incorrectly linked some households or missed ones that should have been linked which may have affected the results. As well, it is not possible to tell whether the family members living in the 'immediate' or 'expanded' household have any influence on the individual's life; however, life in this rural Malawian location tends to be quite communal and family ties and economic obligations strong so it is likely that they do. Equally, family living further away may have an influence: for example, in Malawi older siblings have been shown to positively influence schooling success (Trinitapoli et al., 2014). Techniques have been developed to collect data on social/family networks which might be considered a gold standard (Widmer et al., 2013), however this analysis has shown that at least some attempt at estimating effects beyond the 'immediate' household is possible using secondary data.

Only participants with both parent id-links were used in the LCA so it is possible that other classes exist which could not be detected. Equally, even if an adolescent had both parent identifiers, if a household member did not, it would not be possible to identify them. Households with poor identification of relationships maybe more likely to be unstable (i.e., participants may not have been involved in the HDSS long enough to gather sufficient data) and thus may have different compositions than more stable households. A study on household composition using LCA on historical records found that excluding missing data did cause bias and different results were found when a method allowing for missing data was used (Liao, 2004). However, that study was examining a much simpler outcome, and as few participants were assigned to the 'other' category in the present analysis it was decided not to explore this any further. By using reported biological relationships only, there is a risk of misclassifying people who have other people fulfilling these roles: i.e., a person brought up by their aunt may consider that they live with their 'mother' and 'father' but would not be classified as such in this analysis.

Household compositions are not static and it has been shown in other contexts that changes and instability in household composition has an effect on outcomes (Perkins, 2019). Similar LCA techniques to examine household composition trajectories have been used previously (Huffman et al., 2019a; Mitchell, 2013) which was beyond the scope of this paper but would be a useful future step with these longitudinal data.

The data-driven variables have more categories than the basic and 'traditional' one, the 'immediate' household having 8 classes and the 'expanded' household 11. While this allowed for more nuanced understanding of the data, it should be borne in mind that increasing number of categories may potentially lead to over-fitting of regression models and loss of statistical power to detect effects as the number of degrees of freedom increase. Using categories that are highly specific to the dataset used also makes comparison with other settings more difficult.

Finally, this was a mostly descriptive analysis which involves judgements about how similar or different household definitions are by visually inspecting the distribution of data. Interpretations have been put on this analysis, in terms of how similar/different alternative household definitions are and whether alternative definitions are useful or not in different types of analysis, but different researchers may come up with different interpretations because of the subjectivity involved. This exercise is nevertheless useful because it provides an empirical demonstration of how complex household structure can be, and will allow other researchers to draw their own conclusions about whether they think it might be valuable to use similar methods in their own study contexts.

Conclusions and recommendations

LCA provided a robust approach to better understanding the extent of variation in the data and to guide generation of context appropriate household membership and composition definitions in a location where in-depth ethnographic research was not available. It is recognised that using context-specific household definitions may not be appropriate for many analyses (i.e. if comparison with other settings or analyses is required) and that increasing number of categories used may result in over-fitting of regression models and loss of statistical power. However, where it is appropriate, to benefit from this data-driven method, detailed data on relationships between household members are required and collection of these data should be prioritised when designing population-based studies focused on better understanding the impact of household dynamics. If this is not possible, follow-up questions to more standard questions might be helpful: if 'relationship to household head' is asked, 'relationship to spouse of household head' might be a relatively simple way

to provide much more useful data (i.e. now being able to differentiate between children living with their father and mother or father and step-mother, or understanding that someone living in the household of their in-law is actually living with their sister and her husband). Working in partnership with local researchers and local participants prior to data collection may help development of meaningful definitions or questions to ask.

Researchers examining secondary data might assess whether further household membership classification can be carried out, for example if GPS data are available as in the present analysis. At least there should be some awareness of how households might interact with each other in the context and findings interpreted accordingly. If context-specific variables are not available or appropriate for an analysis, analysts should carefully consider whether what is available has enough meaning in their context to draw useful conclusions.

Acknowledgements

Thank you to all participants and staff of the Karonga HDSS, past and present. Particular thanks are due to Gertrude Longwe and Elizabeth Ndoive who provided valuable insight into local practice and customs. The Karonga HDSS is supported by the Wellcome Trust.

Ethical approval

The Karonga HDSS study has been approved by the Malawian National Health Science Research Committee and the London School of Hygiene and Tropical Medicine Ethics Committee. Informed written consent to participate in the HDSS is given by each household head, which may be rescinded at any time for any reason.

Data availability statement

The data that support the findings of this study are available on reasonable request from the Malawi Epidemiology and Intervention Research Unit (contact the corresponding author in the first instance via info@meiru.mw quoting the paper title). The data are not publicly available as the detailed information about household membership makes it impossible to ensure each individual's anonymity and privacy.

Funding statement

The Karonga Health and Demographic Surveillance Site is supported by The Wellcome Trust, current grant number: 217073.

7. Investigating use of sequence analysis to assess changes in transition to adulthood over time using HDSS data

7.1 Abstract

Background Sequence analysis is a technique which can be used to assess and group individuals' life-course sequences, and has previously been shown to be useful for studying the transition to adulthood. Health and Demographic Surveillance Sites (HDSS) are long-running observational cohorts which regularly collect data that can be used to assign different markers of adulthood (i.e. schooling, living arrangements, marital status, child-bearing). They are potentially useful sources of data to study the transition to adulthood in low and middle income countries, where such data are sparse. However, HDSSs are open geographical cohorts, meaning that once an individual has left the area, no further information is gathered. This reduces the number of people with enough data to create sequences, and also is likely to introduce bias into any estimates produced, as the reasons that adolescents leave the area are likely to be related to the transition to adulthood (i.e. for education or marriage). This analysis aims to assess the utility of sequence analysis to study the transition to adulthood using HDSS data, with a focus on the effect of missing data caused by migrations.

Methods This study used 3 approaches to assess the suitability of sequence analysis for studying the transition to adulthood using HDSS data from rural Northern Malawi. The first approach included women who were present in the HDSS at age 15 and for at least 24 of the subsequent 28 quarters. Multi-channel sequence analysis was used to assess how the timing of transitions in schooling, leaving home, marrying and having children varied or coincided. The second approach also included women who were present in the HDSS at age 15, but did not exclude anyone based on subsequent presence; results from single-channel sequence analysis of marital status were used to assess how including out-migrants might change conclusions, compared to using only a sample of people with more complete sequences. Finally, the third approach drew upon the observations from the first 2, to demonstrate the utility of the technique for in-depth data exploration to help guide later analyses.

Results and conclusions This initial investigation into the use of sequence analysis to study life course transitions using HDSS data has demonstrated both strengths and weaknesses. Although caution is required, as well as a careful understanding of potentially

biases due to migration of participants, it can be used to draw conclusions regarding the transition to adulthood and whether it has changed over time. The results suggest that for many women in the study the transition to adulthood experienced is quite traditional: leaving home to marry and rapidly have children; but that there is some evidence that there has been some change over time, with marrying early becoming less likely and remaining in school and staying in school for longer becoming more common. Sequence analysis can certainly be used as an exploratory tool to generate ideas for research questions, and in particular to aid in visualisation of available data: for HDSS data this is useful as movement in and out of the area is so common.

7.2 Introduction

7.2.1. Importance of studying the transition to adulthood

Adolescence and young adulthood are important stages of development; events and decisions made during these periods can have far-reaching effects into later life (Coall et al., 2016; Day et al., 2015). Most sub-Saharan African countries have young populations, with large numbers of adolescents and young adults. While this ‘demographic dividend’ of a high proportion of young, potentially productive individuals may bring significant benefits to these countries, many such countries are experiencing a period of change (Lutz et al., 2019). With increasing urbanisation and access to schooling and the internet, young people now may expect, and have, very different lives to their parents. Understanding the changing lives of young people is important for ensuring individuals experience healthy and successful transitions to adulthood and for understanding how best to ensure the demographic dividend delivers on its potential for good.

In the scientific community adolescence has been defined as between 10 and 19 years, however, recently there have been calls to extend this to 24 years due to biological (the brain continues to develop at this age (Blakemore and Choudhury, 2006)) and social (traditional ‘adulthood’ events are tending to happen later globally) reasons (Sawyer et al., 2018). There are multiple drivers of delayed adulthood in sub-Saharan Africa. Some changes have occurred through governmental/non-governmental campaigns, such as to improve access to education and reduce child marriage (Daniel, 2017). Changing economic conditions also contribute, as labour markets expand and provide access to jobs beyond agriculture. There is some evidence that young people’s definition of adulthood is changing, with more of a focus on education and employment, as more traditional indicators such as marriage and children become increasingly delayed (Day and Evans, 2015). In many

countries, however, the economic situation (i.e. lack of availability of secondary education, high housing costs and lack of reliable employment) has made it hard for young people to take the traditional steps towards adulthood. This has been termed ‘waithood’: an unwanted period of waiting to become an adult (Hownana, 2012). Paradoxically, while ‘social adulthood’ is being delayed, biological adulthood is occurring earlier as improved nutrition is associated with earlier puberty (Bellis et al., 2006). This increasing disjoint between social and biological adulthood has the potential for social effects, such as increasing sex outside of marriage (Juárez and Gayet, 2014) resulting in increased single parenthood or sexually transmitted infections.

Many studies in sub-Saharan Africa have examined ‘early’ sexual debut, school-leaving, pregnancy and marriage (Delprato et al., 2015; Odimegwu and Mkwananzi, 2016; Yakubu and Salisu, 2018), and a few have also looked at leaving home (Chae et al., 2016). Most studies have assessed each event singly by calculating the proportion who have experienced the events by a certain age. This kind of data can be collected in cross-sectional surveys, are relatively easy to collect and can be very useful for understanding the current situation. More sophisticated studies have collected retrospective data on age/date of events, and some have tried to look at the events in the context of each other by examining the order and timing of events (Beguy et al., 2011; Biddecom and Bakilana, 2003). A recent large analysis of 69 low and middle income countries examine the timing of union formation and child-bearing using sequence analysis (see also below) (Pesando et al., 2021). Such retrospective studies can be subject to recall bias (Mensch et al., 2014) and require greater skill on the part of interviewers to help respondents to report accurate data, but have the benefit of producing results quicker than prospective surveys. These may produce more accurate data, but may also be subject to biases due to selective loss to follow-up (which may be worse in a young population). Few studies in sub-Saharan Africa have examined different trajectories to adulthood people can take (e.g. Goldberg 2013), though it has been shown to be associated with future health (Bennett and Waterhouse, 2018) and HIV infection (Kreniske et al., 2019).

7.2.2. Sequence analysis and its use in studying the transition to adulthood

Sequence analysis is a technique which uses various methods to indicate how similar or dissimilar text strings are from 1 standard string, or from each other. It has been used extensively in genomics for assessing differences in stretches of DNA, and has been used increasingly in life course research over the last few decades. While in genomics each letter in the text strings represents a DNA base pair, in life course research they represent a state that a person experiences, for example marital status. A sequence of NNNNMMMDDDMM

would represent someone spending 4 periods never married, 3 married, 3 divorced and then 2 married again. The period of time that each letter represents varies across analyses, i.e. it could be yearly, quarterly, monthly or even daily. The past, present and future of the use of sequence analysis for life course research has been helpfully summarised in a 2022 review (Liao et al., 2022). In brief: the use of sequence analysis for this kind of research has increased since 2000, and especially since 2010. This explosion of usage lead to multiple debates over techniques and best practices (see also methods section below), some of which may be classed as resolved and some which are still under development (i.e. how to deal with missing data). Sequence analysis has been used to study the transition to adulthood with success with data from Europe (Oris and Ritschard, 2014; Schwanitz, 2017) and more notably for this analysis, in a cross-cultural study using data from 69 low and middle income countries which found the technique useful to generate clusters which enabled comparison in transition to adulthood by region and over time (Pesando et al., 2021). Sequence analysis is not the only technique that can be used to look at transitions over time: a study looking at changes in partnership status over the life-course compared it with latent class growth models and multi-state event history models and found relatively similar conclusions with all three, though the outputs are quite different so it is difficult to compare directly. They found strengths and limitations with all 3 methods and produced a useful comparison table to help choose the method: for example both sequence analysis and latent class growth models can be used to classify individuals but sequence analysis is less computationally intensive (Mikolai and Lyons-Amos, 2017).

7.2.3. Health and Demographic Surveillance Sites and their use for studying transitions

Health and demographic surveillance sites (HDSS) exist across the developing world and function to allow in-depth demographic analyses in locations without robust routine vital statistics collection (Ye et al., 2012). HDSSs routinely collect data on demographic events: births, deaths and migrations, and most tend to also capture information on other socio-demographic factors including schooling, occupation and marital status. These data can cover many typical transitions that adolescents experience, i.e. leaving school, leaving home, getting married and having a child and as these data are collected prospectively the dates and order of the events should be fairly robust, making HDSS data a potentially good source of data to examine transition to adulthood. However, there are also potential limitations: HDSSs are open geographical cohorts where the population in one particular area is under surveillance, but participants are not followed up if they leave the area. While many adolescents may stay living in the area until they are adults, many may leave or move

into the area for schooling, employment or marriage (i.e. due to the adulthood transitions), and this more mobile group may be different compared to those that stay.

There are some examples of HDSS data being used to study transitions in a relatively simple way, i.e. a Thai study looking at the transition to and from smoking from one survey round to the next (Duc et al., 2016), or Malawi data used to manually reduce data on transitions between height for age categories into groups, never stunted, decliners, improvers and always stunted (Sunny et al., 2018). A more complex analysis of South African data looked at multiple transitions on the HIV care cascade, however the transitions were treated as independent (Chang et al., 2018; Haber et al., 2017). Analyses using HDSS data using data linked across time for individuals in a more complex way are less common, however 3 were found. The first, from Zimbabwe, is most relevant to the present analysis as sequences of sexual debut, union formation and child-bearing were created from up to 5 HDSS rounds, grouped and assessed for association with birth cohort and HIV status. This analysis, however, did not used sequence analysis and only generated the sequences based on the order of events. Participants also needed to have experienced at least 2 transitions to be included, and imputation was used for those with incomplete sequences, so people who did not transition were not included (Del Fava et al., 2016). The two other analyses both use sequence analysis, but in quite different ways: the first, from South Africa used dates of birth for live births to construct a sequence to represent the 1000 day period of pregnancy and breast-feeding, the sequence string in this case was a variable relating to HIV/ART care (Etoori et al., 2021). The second analysis, also from South Africa, used sequence analysis to assess participation dynamics across 9 rounds of HIV testing within the HDSS. This analysis only kept people in the analysis if they were present in the HDSS for at least 7 of the 9 rounds, however as the aim was to look at long-term participation in studies so dropping migrants would not bias the conclusions (Larmarange et al., 2015).

The aim of this analysis is to use existing prospectively collected data from the Karonga HDSS in Northern Malawi to, 1. Describe transitions to adulthood and assess whether there is evidence for change over time and 2. Evaluate whether sequence analysis can be usefully applied to HDSS data, in particular, to examine individual trajectories.

7.3. Methods

7.3.1. Data source

Data come from southern Karonga district in Northern Malawi. This is a predominantly rural, subsistence farming and fishing area. The main ethnic group living here are Tumbuka, who since the 19th century have followed patrilineal and patrilocal customs (Vail, 1989): women tend to move to their husband's village when they marry. Land is held by men, and a father will assign land to his married sons (Mutangdura, 2004). Both men and women tend to stay living with their parents/guardian until they marry, though young men may decide to move out earlier to get some freedom, this is not very culturally acceptable (personal communication). Marriage is traditionally negotiated between the groom's paternal uncle and the bride's father and paternal aunts and uncles. If accepted, the bride price (lobola) is agreed upon (Bertrand-Dansereau and Clark, 2016). Anecdotally, elopement, where the bride moves with the groom without negotiation or payment of lobola, is becoming more common due to increase in cost of weddings and lobola (personal communication). In the event of divorce or even paternal death, children that are old enough to be away from their mother may be required to live with their father's family (Malawi Human Rights Commission, 2006; Mwambene, 2012) if the lobola was paid. Marrying and having children is the norm, fertility rates are relatively high but decreasing (McLean et al., 2017). Polygyny is widespread: at the end of 2016 about 15% of households in the HDSS were headed by men with multiple wives. The area has been affected by the AIDS pandemic, though not as greatly as some areas in the country: HIV prevalence was estimated to be 9% in women and 7% in men in 2009 (Floyd et al., 2013). ART became widely available in the area over the period of analysis which contributed to reduced mortality rates and increased life expectancy (Price et al., 2017).

The Karonga Health and Demographic Surveillance System (HDSS) was set up between 2002 and 2004 in a 150m^2 area in the south of the Karonga district (Crampin et al., 2012). All households were surveyed in the initial census and information gathered on household and individuals. The area remains under continuous follow-up: births and deaths are collected on a monthly basis and migrations in, out and within the area, on an annual basis. In-migrants who are returning to the area are consistently and reliably linked back to their original identification number, even if they left the area for a long time and/or moved to a different household. Households are asked to identify a head (who must be resident) and report if/when this changes. Regular surveys gather further information on each household and individual (including their own and their parents' education status), and unique

household and person identifiers allow linkage of all data collected at the site. For all individuals the identification number of their mother and father is recorded, even if the parent is not an HDSS member. All participants over the age of 12 are regularly asked to report their marital status and report past and current spouses, who are also linked to existing personal identifiers (or assigned new ones). For this analysis, data from 1 January 2004 to 31 December 2016 were used.

7.3.2. Data management

The HDSS data are longitudinal: each person's time in the area is arranged as episodes of when they started and ended living in a particular household. This episode dataset was reduced to a quarterly snapshots dataset where the data were reduced to 4 records per person per year (taken at the mid-point of each quarter). Information for other variables are collected at various different times and each snapshot record was categorised for each variable as described below.

Markers of adulthood variables:

Schooling: participants (or close informants) are asked (usually annually) about their schooling, whether they are currently attending and what grade, and if not currently attending what is the highest grade they attended. They are also asked the age or year they left school if they have left. Using this information, each person's schooling history can be constructed, including gaps where they left school and returned. This variable contains categories: never attended, currently in primary, currently in secondary, left primary, left secondary.

Leaving home: using the parental/spouses identifier links, the relationship of the household head to the index person is calculated. This is categorised into parents (including step-parent), other relative or non-relative and self/spouse or in-law.

Marriage: participants (or close informants) are regularly asked (usually annually) to report their current marital status, the start date of their current and previous marriages, and the age/year they first got married. Using the above, a person's marital history can be constructed with the following categories: never married, married, divorced/widowed.

Children: as stated above, mother and father identifiers have been recorded for all individuals continuously during the HDSS. Birth dates or estimates are also recorded for all participants. These were used to classify participants as having no children, one child, or 2

or more children. Unlike the 3 above markers this one is not reversible – there was no attempt to capture whether the children were still living.

When a participant is outside of the HDSS area, no information is gathered, however by using data on the reason for moving, gathered when the migration is registered, plus assessing whether other household members moved from or to the same place as the index (described in detail in McLean et al. 2023) it is possible to assign statuses to time spent living outside of the study area. If a person moved independently and gave the reason for moving as marriage or divorce, then the states of 'migrated for marriage' and 'migrated for divorce' were assigned respectively. This was only applied for people moving independently, as the reasons for move are not necessarily related to the index (i.e. a girl moving because her parents got divorced may have the reason of 'divorce' as well). Other time outside of the area was assigned as 'other' reasons: this will be a mix of accompanied moves for marriage and divorce and independent and accompanied moves for reasons of education, work and other. After experimenting with creating migration categories for the other reasons it was decided that the complexity added did not meaningfully improve the analysis.

7.3.3. Sequence analysis

Sequence analysis was carried out using the TraMineR package in R (Gabadinho et al., 2011). This creates a dissimilarity matrix between all pairs of sequences. There are multiple methods available to create this dissimilarity matrix, all using different ways to assign the 'cost' of changing one sequence into another (through substitution, insertion or deletion of sequence states). A 2016 review of the methods listed 18 and assessed them through simulation studies. They did not find one single method that was superior in all cases, however, made useful recommendations of which to use depending on whether the timing, duration or order of events/states was most important (Studer and Ritschard, 2016). This review was used to guide the method used for each sequence analysis carried out (see below). The dissimilarity matrices were reduced into clusters using the mclust R package (Scrucca et al., 2016) which produced a dendrogram of relatedness between all the sequences. Again, there are multiple clustering algorithms which may be used. The Wards hierarchical method is commonly used; these methods start with each observation as one cluster and at each stage the most similar (according to the results from the dissimilarity matrix) clusters are combined, until eventually all the observations are included in one large cluster. Wards algorithm has been criticized for being biased towards producing clusters of similar sizes and being sensitive to outliers (Lesnard, 2006). The same reviewer suggested other techniques to use, however for this analysis these other methods produced unhelpful results (i.e. clusters only containing one participant) so Wards was used throughout.

For each analysis, a range of cluster numbers were produced and each displayed on a sequence index graph which allow for visual assessment of the characteristics of each cluster: these figures present each sequence within each group as a line, with different colours representing the different sequence states (Brzinsky-Fay, 2014). In addition, several statistical measures of the quality of the different cluster solutions were calculated using the WeightCluster (Studer, 2013) package in R. These give an indication of the number of clusters that are optimal. They tended not to be particularly helpful, some cluster statistics usually favour the highest number possibles, others the lowest number, so while they were used to guide the process, the following process was used to select the 'optimal' cluster solution. Each single decrease in number of clusters is the result of two clusters being merged, so the smaller cluster solution was compared by eye to the next largest one to identify which clusters had been merged: if the 2 merged clusters were considered to be different enough from each other, and the 2 groups included enough participants (trying to keep at least 100 participants per cluster), the smaller cluster solution was rejected and the larger one compared to the next one. Once it was felt that the 2 merged clusters were not different enough, or were too small for meaningful analysis, the larger cluster solution was rejected and the smaller one kept. The term 'different enough' is of course subjective and did depend on the individual analysis somewhat, for example in the initial exploratory analysis difference in timing were felt to be of less interest than in later analyses (described more in the results section).

3 approaches using the HDSS data using sequence analysis were carried out, these represent an iterative process and refining of the technique to balance with the strengths and limitations of the HDSS data (this process is described in the results). For all approaches, for simplicity, only female participants were used as firstly more research has been carried out on women making the results easier to interpret in the context of the existing literature and, in this area of Malawi, women are more likely to move due to marriage so would represent a 'harder' test of the technique and data to answer the question.

1. Approach 1 utilised multi-channel sequence analysis. In this form of sequence analysis, each participant has a separate sequence for each of 2 or more domains. The resulting dissimilarity matrices are combined into one before cluster analysis is carried out, however the sequence index figures display each domain separately. Multichannel sequence analysis has advantages over just using one variable as it allows the observation of the inter-connectedness of life experiences. In some cases it may be used to understand how other life processes or events interact with or impact the main outcome under investigation, or the

interactions of all factors may be the goal of the analysis (Pollock, 2007). My hypothesis was that all four factors contribute to the transition to adulthood so multi-channel sequence analysis seemed like a good starting point for exploratory analysis.

Participants were included if they were present in the HDSS at age 15 and then for 24 of the subsequent 28 quarters (7 years): the majority of women will experience the transitions in this period, while a longer period would capture information on earlier and later transitions, it would also reduce the number of sequences that could be used for the analysis. Hamming distance method was used, with default substitution costs. This method is recommended when the focus is on the timing of the events as it only allows for substitutions between states, not insertions or deletions (Studer and Ritschard, 2016), and this was considered appropriate as one of the aims was to see how the 4 variables interact over time. Wards clustering algorithm was used. Descriptive statistics regarding the average length of time spent in each state, and the average age of transition were carried out along with comparison of the proportion in each cluster in the early birth cohort (up to 1992) compared to the later one (1993 onwards). Additionally, the representativeness of the sample was assessed with 3 cross-sectional analyses comparing the 4 adulthood markers in included participants with those excluded (due to being present for less than 24 of 28 quarters after the age of 15) at the ages of 14 (before the start of analysis time), 17 and 20. Individuals may appear in more than one of the age-specific analyses. Chi^2 tests were carried out for each level of each marker (using binary dummy variables) to test for the association between inclusion and being in each marker category.

2. Approach 2 aimed to further assess the impact of losing data on participants who leave the area during the sequence period. It used single-channel sequence analysis with the marital status variable from approach one, using the full dataset of women who were present in the HDSS for at least 1 quarter at age 15 and were born early enough that they could have been present for the following 28 quarters (born before 1997). Periods spent outside of the HDSS were categorised as ‘migrated for marriage’, ‘migrated for divorce’ or ‘migrated for other reasons’ as described above. For consistency with approach 1, Hamming distance methods was used but with user-defined substitution matrices which set the substitution costs between ‘married’ and ‘migrated for marriage’, and ‘divorced’ and ‘migrated for divorce’ as 0.2, while cost for all other substitutions were 1. Setting substitution costs is always an arbitrary process: in this case the very small relative cost for these states means that they would be treated almost as if they were the same state (which would have a substitution cost of 0). Wards clustering algorithm was used. After comparison of the resulting solutions, the sequences were manually categorised into clusters (guided by the clusters produced by the

algorithm) and the percent change in categories from the earlier to the later birth cohort compared with the results from the dataset only including participants who were present for 24 of 28 quarters.

3. Approach 3 drew upon the findings from approach 1 and 2 to demonstrate the utility of sequence analysis for initial exploration of a specific question. Participants were included if they had had a child while unmarried before the age of 18, by the end of 2013 (to enable at least 4 years of follow-up time before the end of analysis time [end 2017]). The marital status variable from approach 2 was used (including the different categories for migration), and the sequence length of was 20 quarters (5 years). The sequencing was used simply as a visualisation tool for initial assessment of whether the trajectories were different in 2 calendar eras.

7.4. Results

7.4.1. Approach 1: Initial exploration of transitions using multi-channel sequence analysis

1665 adolescents born before 1997 were present in the HDSS at age 15, and for 24 of the subsequent 28 quarters so were included in the first analysis. The four-cluster solution was chosen for the multi-channel sequence analysis; this decision was based solely on the examination of the sequence index plots. Some of the cluster statistics recommended the 8-cluster solution (table 7.1): the additional clusters were mostly splits on timing (i.e. splitting the first group into those who are already married and those who marry in the first few quarters [figures not shown]), but the overall conclusions did not change so it was decided to use the simplest solution. The sequence index plots (figures plotting each individual sequence) for each cluster and each adulthood marker variable are shown in figure 7.1, and descriptive statistics showing the number of participants experiencing each state, the median number of quarters and the median age of first transition for each marker and cluster are shown in table 7.2.

Table 7.1. Cluster statistics for first 2 sequence analyses. For each approach and statistic, the 'best' and 'second best' number of clusters, along with the cluster statistic is displayed. The actual number of clusters chosen is shown in the table heading.

	1. Multichannel (at least 24 of 28 quarters) *Selected solution: 4 clusters*				2. Marital status (full data) *Selected solution: 9 clusters*			
	Best		Second best		Best		Second best	
	clusters	stat	clusters	stat	clusters	stat	clusters	stat
PBC	2	0.648	6	0.611	5	0.653	3	0.640
HG	8	0.876	9	0.874	10	0.890	9	0.889
HC	8	0.070	9	0.072	9	0.083	10	0.083
HGDS	8	0.874	9	0.872	10	0.886	9	0.885
ASW	2	0.398	6	0.287	3	0.398	5	0.352
ASWW	2	0.398	6	0.398	3	0.398	5	0.353
CH	2	612.7	3	412.7	3	1114.0	4	917.8
CHSQ	2	1392.8	3	932.7	3	2389.0	4	0.6
R2	10	0.522	9	0.509	10	0.597	9	0.580
R2SQ	10	0.748	9	0.738	10	0.793	9	0.780

PBC (Point Biserial Correlation): Measure of the capacity of the clustering to reproduce the distances ([−1;1] Max); HG (Hubert's Gamma): Measure of the capacity of the clustering to reproduce the distances (order of magnitude) ([−1;1] Max); HC (Hubert's C): Gap between the partition obtained and the best partition theoretically possible with this number of groups and these distances ([0;1] Min); HGSD (Hubert's Somers' D): Measure of the capacity of the clustering to reproduce the distances (order of magnitude) taking into account ties in distances ([−1;1] Max); ASW (Average Silhouette Width): Coherence of assignments. High coherence indicates high between-group distances and strong within-group homogeneity ([−1;1] Max); ASWW (Average Silhouette Width (weighted)): As previous, for floating point weights ([−1;1] Max); CH (Calinski-Harabasz index): Pseudo F computed from the distances ([0;∞] Max); CHsq: As previous, but using squared distances ([0;∞] Max); R2 (Pseudo R2): Share of the discrepancy explained by the clustering solution (only to compare partitions with identical number of groups) ([0;1] Max); R2sq: As previous, but using squared distances ([0;1] Max)

The four clusters are described below:

Cluster 1 (N=743, 44.6%): This group leaves primary school, home and marries, the earliest in the period, and transitions to one and then more children rapidly. There appears to be a small section, however, who leave primary school, have a child, but do not marry (figure 7.1). Almost all quarters were spent as left primary school, married and living with spouse. The mean age of leaving school, leaving home and getting married was 16 or 17 and having a child was 17 (table 7.2). The proportion in this cluster decreased between the 2 birth cohorts (48.4% to 41.2%).

Cluster 2 (N=212, 12.7%): Members of this cluster attend secondary school but leave, marry and have children quite early in the period (though slightly later than in cluster 1) (figure 7.1). On average people in this group spend about 15 quarters married (compared to 19 in group 1), and the mean age of leaving schooling, home and marrying was 17 or 18 and for having a child was 18 (table 7.2). The proportion in this cluster was higher for the earlier birth cohort (14.0%) compared to the later one (11.6%).

Cluster 3 (N=235, 14.1%): This group is characterised by living in a non-parental, non-marital household, most attend secondary school and marriage tends to happen later than cluster 1 & 2 and only about half are married by the end of the period (figure 7.1). The group spends on average 8 quarters in primary school and 11 in secondary school, and those that marry do so at about 19 years old (table 7.2). The proportion was similar in the 2 birth cohorts (13.9% and 14.3%).

Cluster 4 (N=475, 28.5%): This group marry the latest, or do not marry by the end of the period, most live with parents and attend secondary school (figure 7.1). This group spend an average of 7 quarters in primary school and 16 in secondary school, and those who marry do so at about 20 years old (table 7.2). The proportion in this group increased between the 2 birth cohorts, from 23.6% to 33.0%.

The main differences in the clusters are related to timing (earlier or later), whether secondary school is attended and whether living in a non-parental, non-marital house. However, all clusters include a mix of the 2 non-independent households (with parents and other), and there is no evidence that participants were moving between these 2 types: participants start in one or the other and transition to in-law/spouse/self-headed households (and back again possibly). While common, divorce does not have an effect on clusters, and the timing of the transitions tends to be similar: i.e. a group leaving school at a particular time will leave home and marry at a similar time, then have children with a slight delay. There are no groups where only one or 2 transitions are experienced, however it is possible to see that there are some women who have a child but do not marry who are scattered across multiple groups.

The 1665 included in the analysis above represent a subset (47%) of the 3548 people who were present at age 15 and born before 1997. Table 7.3 shows the number and proportion of 14, 17 and 20-year olds according to whether they were included in the analysis or not, by the 4 adulthood markers, plus the results of Chi^2 tests assessing the association between inclusion and adulthood markers. Being included was associated with difference in all 4

markers for all 3 ages: the trends vary slightly by age, but in general tend towards implying that included participants may be more likely to experience the transition to adulthood faster than those not included: at age 14 (before the sequence period) those included were more likely to be in primary school, however by age 17 the included group were more likely to have left primary school. At age 14 included participants were more likely to be living with parents, and at all ages, included participants were less likely to be living in a non-parental, non-marital household. At age 17 and 20 included participants were more likely to be married but less likely to be divorced. Included participants were more likely to have children at age 17, while at age 20 included participants were less likely to have 1 child, but more likely to have 2 or more children (table 7.3).

Figure 7.1. Sequence index plots for the 4-cluster solution from multi-channel sequencing for the 4 markers: schooling, household head, marital status and child status

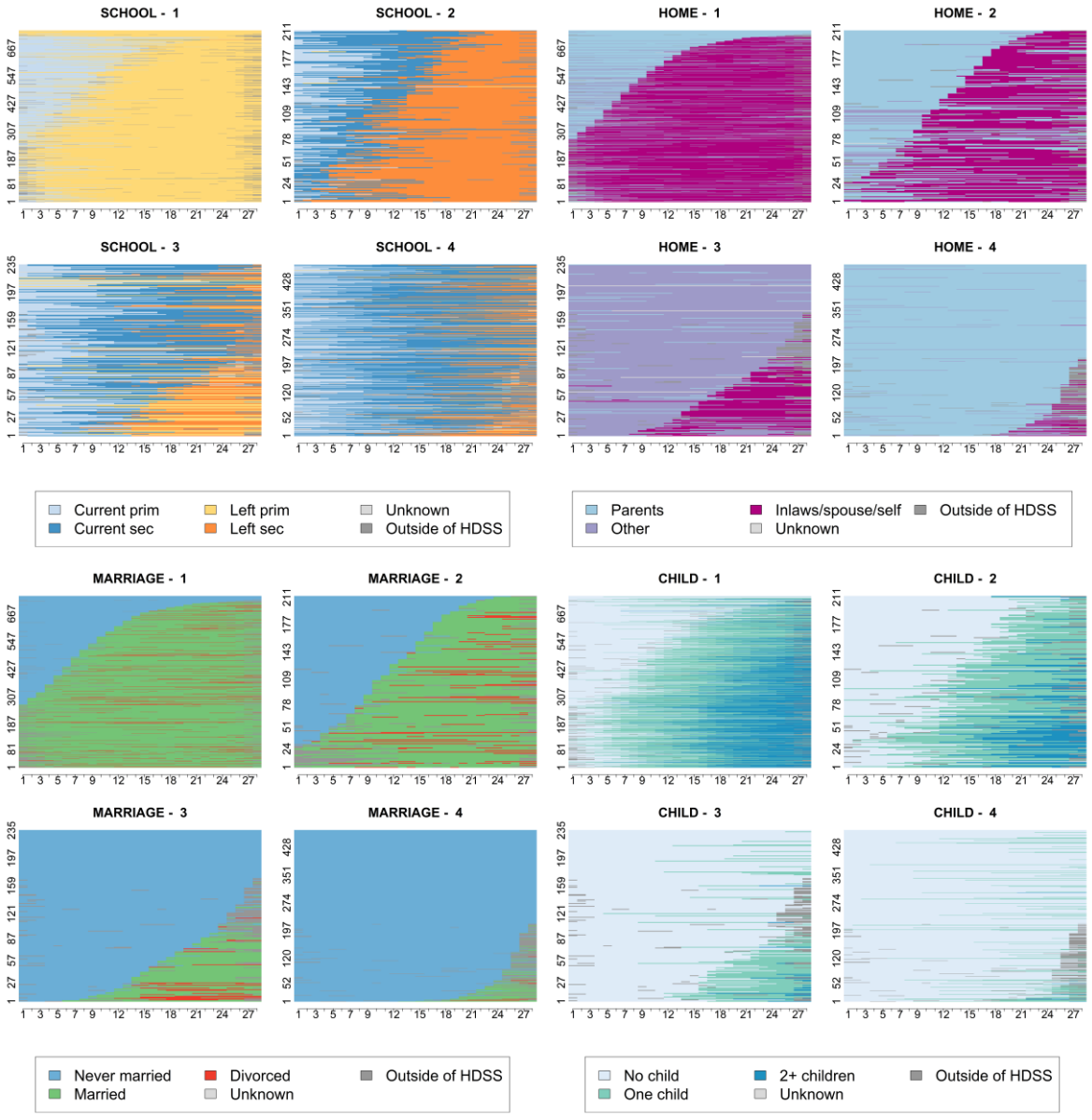


Table 7.2. Descriptive statistics for 4 clusters generated through multi-channel sequence analysis. Overall number (and proportion), and mean (and standard deviation) per person, of quarters spent in each category is shown, along with the mean (and standard deviation) of the youngest age at each transition.

	Early marriage, primary (n=743)		Mid marriage, secondary (n=212)		Late/no marriage, no parents (n=235)		Late/no marriage, parents (n=475)	
	Num (%)	Mean (sd)	Num (%)	Mean (sd)	Num (%)	Mean (sd)	Num (%)	Mean (sd)
Schooling								
Current primary	541 (72.8%)	5.3 (5.0)	132 (62.3%)	3.4 (3.8)	196 (83.4%)	7.8 (6.3)	332 (69.9%)	6.8 (6.7)
Current secondary	2 (0.3%)	0.0 (0.6)	210 (99.1%)	8.2 (5.1)	179 (76.2%)	11.2 (8.2)	436 (91.8%)	15.6 (7.5)
Left primary	742 (99.9%)	21.7 (5.1)	3 (1.4%)	0.1 (0.7)	52 (22.1%)	3.6 (7.3)	34 (7.2%)	0.7 (2.8)
Left secondary	1 (0.1%)	0.0 (0.3)	210 (99.1%)	15.2 (5.2)	108 (46.0%)	4.2 (5.7)	221 (46.5%)	3.9 (5.5)
Unknown	1 (0.1%)	0.0 (0.1)	1 (0.5%)	0.0 (0.1)	15 (6.4%)	0.3 (1.9)	40 (8.4%)	0.2 (0.7)
Outside HDSS	319 (42.9%)	1.0 (1.3)	94 (44.3%)	1.1 (1.4)	107 (45.5%)	1.1 (1.4)	157 (33.1%)	0.8 (1.3)
Age left school		16.1 (1.2)		17.6 (1.3)		18.6 (1.6)		19.2 (1.5)
Household head								
Parent/step-parent	448 (60.3%)	7.1 (8.3)	177 (83.5%)	13.0 (8.0)	40 (17.0%)	1.4 (3.6)	474 (99.8%)	25.9 (3.4)
Other	191 (25.7%)	1.9 (4.1)	53 (25.0%)	1.7 (3.5)	231 (98.3%)	21.6 (6.5)	47 (9.9%)	0.7 (2.5)
Inlaw/spouse/self	695 (93.5%)	17.9 (8.1)	198 (93.4%)	12.2 (6.9)	94 (40.0%)	3.5 (5.3)	68 (14.3%)	0.6 (1.8)
Unknown	15 (2.0%)	0.2 (1.5)	2 (0.9%)	0.0 (0.4)	5 (2.1%)	0.4 (2.8)	0 (0.0%)	0.0 (0.0)
Outside HDSS	319 (42.9%)	1.0 (1.3)	94 (44.3%)	1.1 (1.4)	107 (45.5%)	1.1 (1.4)	157 (33.1%)	0.8 (1.3)
Age left home		16.4 (1.4)		17.6 (1.4)		18.8 (1.4)		20.5 (0.7)

	Early marriage, primary (n=743)		Mid marriage, secondary (n=212)		Late/no marriage, no parents (n=235)		Late/no marriage, parents (n=475)	
	Num (%)	Mean (sd)	Num (%)	Mean (sd)	Num (%)	Mean (sd)	Num (%)	Mean (sd)
Marital status								
Never	470 (63.3%)	5.7 (6.8)	185 (87.3%)	9.8 (6.3)	235 (100.0%)	22.0 (6.3)	475 (100.0%)	26.2 (2.7)
Married	714 (96.1%)	19.1 (7.6)	209 (98.6%)	14.5 (6.3)	101 (43.0%)	3.9 (5.6)	84 (17.7%)	0.8 (2.1)
Divorced	165 (22.2%)	1.7 (4.2)	55 (25.9%)	1.9 (4.1)	22 (9.4%)	0.7 (2.6)	8 (1.7%)	0.1 (0.6)
Unknown	0 (0.0%)	0.0 (0.0)	0 (0.0%)	0.0 (0.0)	0 (0.0%)	0.0 (0.0)	0 (0.0%)	0.0 (0.0)
Outside HDSS	385 (51.8%)	1.5 (2.0)	114 (53.8%)	1.8 (2.5)	115 (48.9%)	1.4 (1.8)	164 (34.5%)	0.9 (1.5)
Age first married		16.1 (1.3)		17.3 (1.4)		18.8 (1.3)		20.3 (0.8)
Children								
No children	716 (96.4%)	10.2 (6.6)	208 (98.1%)	13.8 (6.3)	235 (100.0%)	22.8 (5.4)	474 (99.8%)	25.1 (5.1)
One child	712 (95.8%)	10.5 (4.5)	204 (96.2%)	10.1 (4.4)	115 (48.9%)	3.8 (4.9)	119 (25.1%)	2.0 (4.7)
2 or more children	508 (68.4%)	6.3 (5.9)	84 (39.6%)	3.1 (4.7)	20 (8.5%)	0.3 (1.3)	7 (1.5%)	0.1 (0.6)
Unknown	0 (0.0%)	0.0 (0.0)	0 (0.0%)	0.0 (0.0)	0 (0.0%)	0.0 (0.0)	0 (0.0%)	0.0 (0.0)
Outside HDSS	319 (42.9%)	1.0 (1.3)	94 (44.3%)	1.1 (1.4)	107 (45.5%)	1.1 (1.4)	157 (33.1%)	0.8 (1.3)
Age first child		17.2 (1.5)		18.1 (1.5)		19.4 (1.2)		19.4 (1.7)

Table 7.3. Assessment of the representativeness of the sample included for multi-channel sequence analysis compared to participants present in the HDSS at 15, born before 1997 but not included due to not being present for at least 24 of 28 quarters at 3 ages

	Age 14					Age 17					Age 20				
	Not included		Included			Not included		Included			Not included		Included		
	N	%	N	%	p-value	N	%	N	%	p-value	N	%	N	%	p-value
Schooling															
Curr Prim	1095	72.7	1109	79.9	<0.001	266	24.1	293	17.6	<0.001	4	0.8	15	0.9	0.852
Curr Sec	190	12.6	155	11.2	0.229	345	31.2	571	34.4	0.086	107	21.7	354	21.3	0.833
Left prim	139	9.2	118	8.5	0.491	329	29.8	668	40.2	<0.001	240	48.7	820	49.2	0.825
Left sec	1	0.1	1	0.1	0.954	68	6.2	129	7.8	0.107	124	25.2	460	27.6	0.277
NK	81	5.4	5	0.4	<0.001	97	8.8	1	0.1	<0.001	18	3.7	16	1.0	<0.001
Household head															
Par/step-par	848	56.3	933	67.2	<0.001	520	47.1	758	45.6	0.453	194	39.4	569	34.2	0.035
Other	590	39.2	383	27.6	<0.001	321	29.0	265	15.9	<0.001	96	19.5	182	10.9	<0.001
Inlaw/spouse/self	54	3.6	63	4.5	0.193	256	23.2	630	37.9	<0.001	201	40.8	905	54.4	<0.001
NK	14	0.9	9	0.6	0.395	8	0.7	9	0.5	0.547	2	0.4	9	0.5	0.712
Marital status															
Never	1316	87.4	1242	89.5	0.078	712	64.4	910	54.8	<0.001	153	31.0	587	35.3	0.083
Married	74	4.9	82	5.9	0.237	276	25.0	644	38.7	<0.001	211	42.8	900	54.1	<0.001
Div/wid	48	3.2	13	0.9	<0.001	101	9.1	100	6.0	0.002	129	26.2	177	10.6	<0.001
NK	68	4.5	51	3.7	0.255	16	1.4	8	0.5	0.007	0	0.0	1	0.1	0.586
Children															
No children	1491	99.0	1372	98.8	0.682	853	77.2	1107	66.6	<0.001	172	34.9	596	35.8	0.712
1 child	15	1.0	16	1.2	0.682	232	21.0	506	30.4	<0.001	211	42.8	582	35.0	0.002
2+ children	0	0.0	0	0.0		20	1.8	49	2.9	0.060	110	22.3	487	29.2	0.002

7.4.2. Approach 2: assessing the effect of including missing data periods

As the multi-channel analysis above showed that the 4 transitions tended to show similar patterns in terms of timing, for simplicity only the marriage variable was used for the next approach, to look at the effect of out-migration. All women who were present in the HDSS at age 15 for at least 1 quarter, and born before 1997 were included (n=3548).

The 9-cluster solution was selected: this was the recommended 'best' solution for 1 of the 10 cluster quality statistics, and 'second best' for 4 (table 7.1), and also was felt to produce useful clusters through visual examination. 2 of the clusters were characterised by the 'outside of HDSS for other reasons' state but with slightly different timings: these were rejoined as the lack of information about the reason for leaving made conclusions about both groups the same. The sequence index plot for each cluster is shown in figure 7.2, descriptive statistics in table 7.4, and are described below:

Cluster 1 (N=406, 11.4%): The majority are already married by the beginning of the period (figure 7.2). There were few participants who migrate for marriage in this group (34, 8.4%) (table 7.4).

Cluster 2 (N=500, 14.1%): All have married or left for marriage by the end of the 3rd year of analysis (figure 7.2). The mean age of marriage is 16.5 and 31% migrated for marriage (table 7.4).

Cluster 3 (N=363, 10.2%): Most get married or leave for marriage in the mid-period (figure 7.2). The mean age of marriage was 18.4 and 25% migrated for marriage (table 7.4).

Cluster 4 (N=782, 22.0%): This group either marry later in the period or are not married by the end (50% remain never married) (figure 7.2).

Cluster 5 (N=137, 3.9%): This cluster is characterized by experiencing divorce, participants marry early, or by the mid-point, and most divorce relatively quickly after marriage (figure 7.2). The mean age of marriage was 16.3, and 27% migrated for divorce (table 7.4).

Cluster 6 (N=134, 3.8%): This group tends to marry and leave, but to return, the majority are married by the end (figure 7.2). This groups also includes people who do not leave: only 84% have periods of 'outside of HDSS for other reasons' (table 7.4).

Cluster 7 (N=148, 4.2%): This group also tends to marry and leave the area, however they do not return (figure 7.2), 100% of this group spend some time in the ‘outside of HDSS for other reasons’ category) (table 7.4).

Cluster 8 (N=1078, 30.4%): This group is 2 clusters recombined, as they split into 2 groups based on timing of leaving. The group is characterised by participants leaving the area for other or unknown reasons without marrying (100% experience this state) (table 7.4).

Figure 7.2. Sequence index plots for the 8 clusters (the 9-cluster solution with 2 clusters combined) generated through single-channel sequence analysis including all migrants

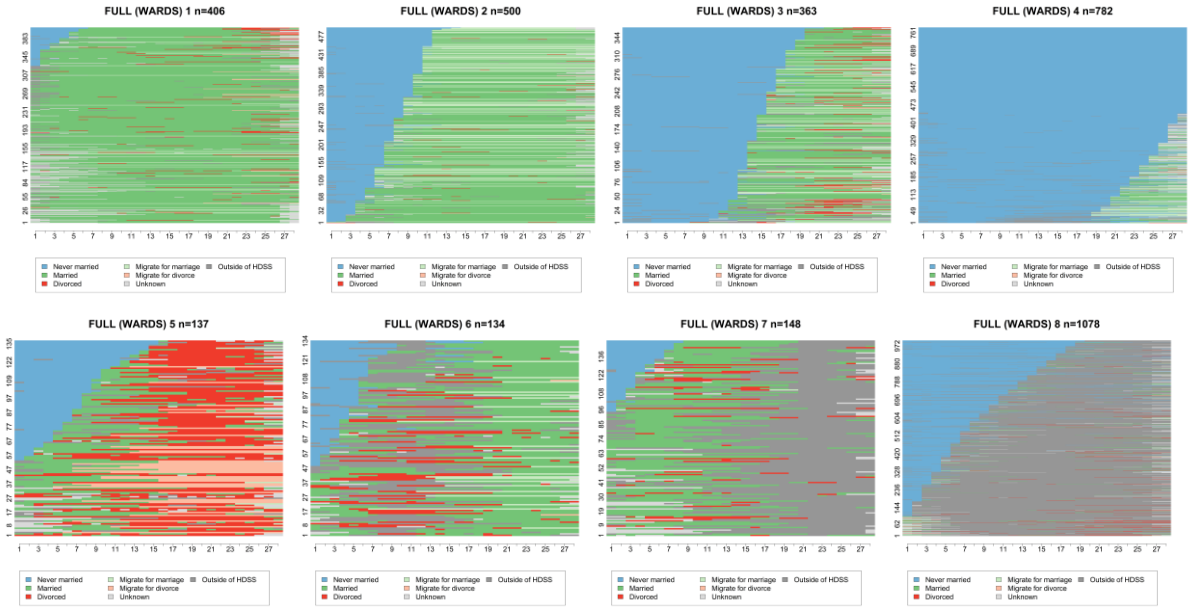


Table 7.4. Descriptive statistics for 8 clusters (the 9-cluster solution with 2 clusters combined) generated through single-channel sequence analysis including all migrants. Overall number (and proportion), and mean (and standard deviation) per person, of quarters spent in each category is shown, along with the mean (and standard deviation) of the youngest age at marriage.

	Earliest marriage (n=406)		Early/mid marriage (n=500)		Mid marriage (n=363)		No/late marriage (n=782)	
	Num (%)	Mean (sd)	Num (%)	Mean (sd)	Num (%)	Mean (sd)	Num (%)	Mean (sd)
Never married	80 (19.7%)	0.5 (1.2)	497 (99.4%)	7.1 (2.7)	363 (100.0%)	14.3 (2.6)	782 (100.0%)	25.1 (3.6)
Married	386 (95.1%)	22.2 (6.9)	351 (70.2%)	13.7 (9.5)	274 (75.5%)	7.8 (5.5)	138 (17.6%)	0.8 (2.0)
Divorced	81 (20.0%)	0.9 (2.1)	39 (7.8%)	0.2 (1.0)	75 (20.7%)	0.9 (2.1)	10 (1.3%)	0.0 (0.3)
Migrate for marriage	34 (8.4%)	1.3 (4.8)	154 (30.8%)	5.8 (8.9)	90 (24.8%)	3.0 (5.5)	84 (10.7%)	0.7 (2.1)
Migrate for divorce	22 (5.4%)	0.3 (1.5)	5 (1.0%)	0.0 (0.2)	6 (1.7%)	0.1 (0.5)	0 (0.0%)	0.0 (0.0)
Unknown	180 (44.3%)	1.8 (2.9)	143 (28.6%)	0.6 (1.1)	82 (22.6%)	0.5 (1.1)	176 (22.5%)	0.6 (1.3)
Outside of HDSS	148 (36.5%)	1.0 (1.7)	93 (18.6%)	0.5 (1.2)	113 (31.1%)	1.4 (2.7)	183 (23.4%)	0.9 (2.2)
Marriage age		15.3 (0.7)		16.5 (0.7)		18.4 (0.7)		20.2 (0.6)

Table 7.4. continued

	Divorce (n=137)		Return (n=134)		Marry then leave (n=148)		Leave/no info (n=1078)	
	Num (%)	Mean (sd)	Num (%)	Mean (sd)	Num (%)	Mean (sd)	Num (%)	Mean (sd)
Never married	84 (61.3%)	4.5 (4.6)	86 (64.2%)	2.9 (2.8)	54 (36.5%)	1.4 (2.2)	977 (90.6%)	8.2 (6.5)
Married	127 (92.7%)	6.6 (4.5)	125 (93.3%)	12.3 (5.8)	136 (91.9%)	9.3 (5.4)	188 (17.4%)	0.6 (1.7)
Divorced	101 (73.7%)	9.5 (7.0)	58 (43.3%)	3.0 (4.3)	47 (31.8%)	1.7 (3.6)	83 (7.7%)	0.4 (1.8)
Migrate for marriage	11 (8.0%)	0.3 (1.0)	23 (17.2%)	1.9 (4.6)	1 (0.7%)	0.0 (0.2)	10 (0.9%)	0.0 (0.3)
Migrate for divorce	37 (27.0%)	4.4 (7.6)	2 (1.5%)	0.1 (0.4)	0 (0.0%)	0.0 (0.0)	1 (0.1%)	0.0 (0.2)
Unknown	53 (38.7%)	1.2 (2.2)	37 (27.6%)	0.7 (1.3)	47 (31.8%)	1.4 (2.9)	291 (27.0%)	0.8 (1.5)
Outside of HDSS	64 (46.7%)	1.6 (2.1)	113 (84.3%)	7.1 (4.7)	148 (100.0%)	14.1 (4.8)	1078 (100.0%)	18.0 (6.5)
Marriage age		16.3 (1.4)		16.5 (1.6)		15.4 (0.7)		17.8 (2.3)

To be able to compare conclusions with the full (born before 1997 and present in the HDSS at age 15 for at least 1 quarter) or restricted (born before 1997, present in the HDSS at age 15 and for at least 24 of the subsequent 28 quarters) groups, the clusters found were used as guidance to create a manually categorised variable with 6 categories: earliest marriage (married or migrated for marriage by quarter 4), early marriage (married or migrated for marriage in quarter 5-12), mid-marriage (quarter 13-24), no or late marriage (quarter 25-28 or never married by the end), ever divorce (ever reported as divorced or migrated for divorce) and left for other reasons. Sequence index figures for the 6 categories using the full and reduced datasets are shown in figure 7.3. The proportion in each category overall and in the 2 birth cohorts, along with percent differences are shown in table 7.5, and the percent differences shown graphically in figure 7.4. For both datasets, between the early and later birth cohort, there is a decrease in all categories except the no/late marriage category which increased. For the first 3 marriage categories (earliest, early and mid) using the restricted dataset suggested a larger decrease compared to the full dataset, however the confidence intervals overlap. Similarly, the percent increase in the no/late marriage category is larger

when using the restricted dataset compared to the full dataset, but again the confidence intervals overlap. There was little difference in the percent decrease in the divorced category using the 2 datasets, and while the left for other reasons category decreased with the full dataset, it did not exist with the restricted dataset.

Including the full dataset with categories for leaving for marriage and divorce firstly enables migrants to be categorised similarly to those who stay in the area which hopefully makes the proportions in each group more representative of the population. However, the downside of doing so is that it implies knowledge of the person's state possible long after the confidence has gone, and especially in this case where the state of marriage is reversible: some of the migrated for marriage participants may actually be divorced and therefore in the wrong category. Where the reason for migration is other or unknown the groups show at least 3 different clusters of behaviours (i.e. leaving but returning, leaving after marriage and leaving before marriage), while it may not be possible to draw clear analytical conclusions for these groups, it does help to understand the underlying data in depth. While the dataset using the full data cannot be considered the gold-standard for comparing the percent differences in categories between the eras, the results imply that for some of the categories the results using the restricted dataset may still be valid, although the increase in the largest group does need to be treated with caution.

Figure 7.3. Sequence index plots for the 6 manually created categories, comparing the results from the full dataset (including time spent outside the HDSS) and the sample restricted to those present for at least 24 of 28 quarters

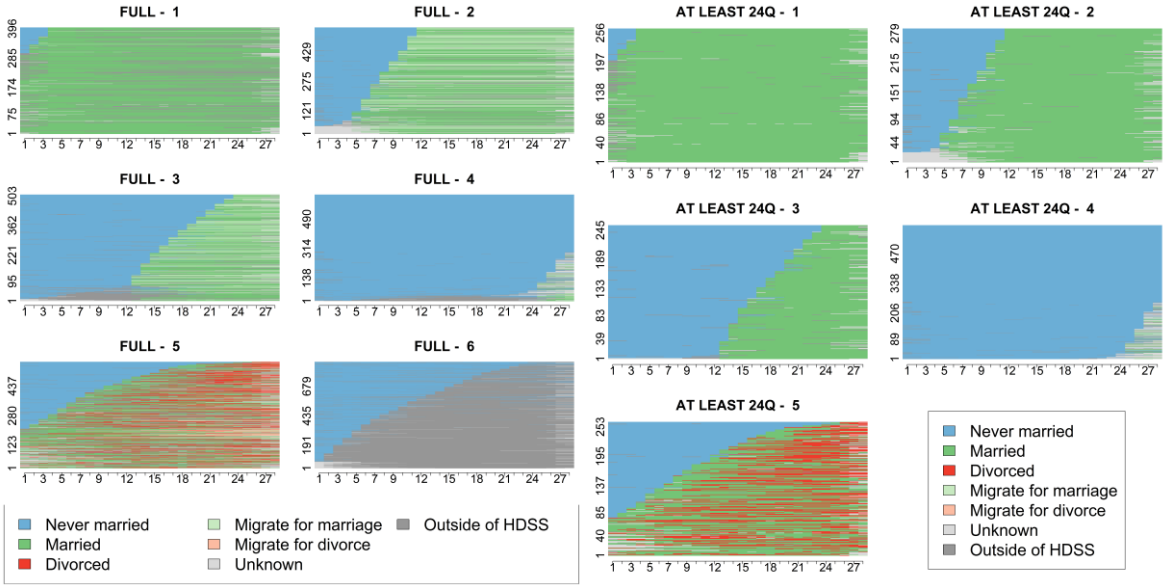


Figure 7.4. Percent differences of manual sequence categories for later birth cohort compared to earlier birth cohort for full dataset and dataset including only those present for at least 24 of 28 quarters.

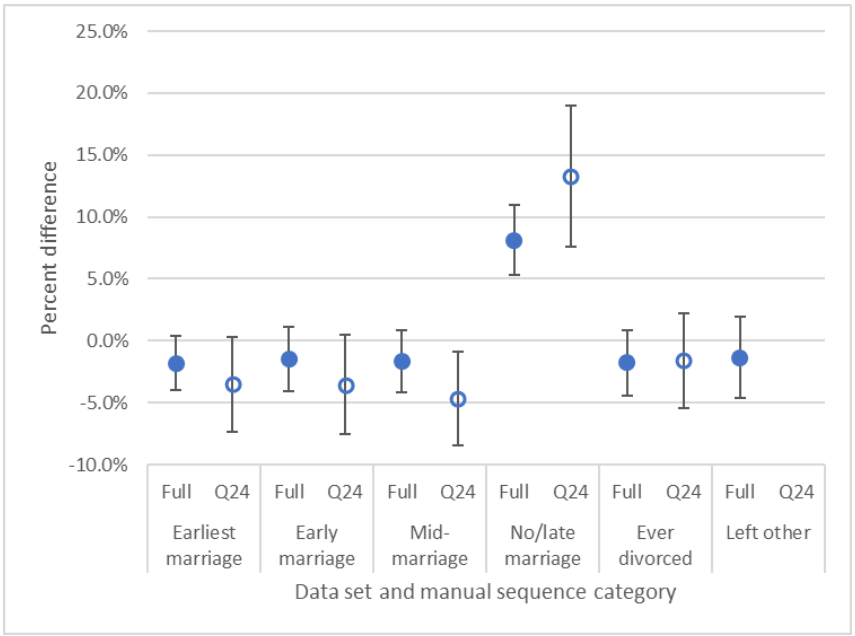


Table 7.5. Comparison in number and percent in manually created sequence categories using full dataset and dataset restricted to those present for at least 24 of 28 quarters

	Total				Birth cohort 1				Birth cohort 2						
	N	%	95% CI		N	%	95% CI		N	%	95% CI		% diff	95% CI	
Full dataset															
Earliest															
marriage	400	11.3%	10.3%	12.4%	219	12.2%	10.7%	13.8%	181	10.4%	9.0%	11.9%	-1.8%	-4.0%	0.4%
Early marriage	557	15.7%	14.5%	16.9%	296	16.4%	14.8%	18.2%	261	14.9%	13.3%	16.7%	-1.5%	-4.1%	1.1%
Mid-marriage	509	14.3%	13.2%	15.5%	273	15.2%	13.5%	16.9%	236	13.5%	11.9%	15.2%	-1.7%	-4.2%	0.8%
No/late marriage	635	17.9%	16.6%	19.2%	250	13.9%	12.3%	15.6%	385	22.0%	20.1%	24.0%	8.1%	5.3%	10.9%
Ever divorced	566	16.0%	14.8%	17.2%	303	16.8%	15.1%	18.6%	263	15.0%	13.4%	16.8%	-1.8%	-4.4%	0.8%
Left other	881	24.8%	23.4%	26.3%	459	25.5%	23.5%	27.6%	422	24.1%	22.2%	26.2%	-1.4%	-4.6%	1.9%
Total	3548				1800				1748						
Present at least 24 quarters															
Earliest															
marriage	264	15.9%	14.1%	17.7%	140	17.7%	15.1%	20.5%	124	14.2%	11.9%	16.7%	-3.5%	-7.4%	0.3%
Early marriage	289	17.4%	15.6%	19.3%	152	19.2%	16.5%	22.1%	137	15.7%	13.3%	18.3%	-3.5%	-7.6%	0.5%
Mid-marriage	254	15.3%	13.6%	17.1%	140	17.7%	15.1%	20.5%	114	13.0%	10.9%	15.5%	-4.7%	-8.4%	-0.9%
No/late marriage	596	35.8%	33.5%	38.2%	228	28.8%	25.7%	32.1%	368	42.1%	38.8%	45.5%	13.3%	7.6%	19.0%
Ever divorced	262	15.7%	14.0%	17.6%	131	16.6%	14.0%	19.3%	131	15.0%	12.7%	17.5%	-1.6%	-5.4%	2.3%
Left other	0				0				0						
Total	1665				791				874						

7.4.3. Approach 3: Sequence analysis for initial exploration of a specific question

As one of the observations of approach one was the inability to separate the rare group of women who had a child without being married, it was decided to use sequencing to examine this group more thoroughly, as an example of a research question that might be posed. Despite reservations about using the migrated for marriage/divorce categories in approach 2, it was felt that they would be useful for an exploratory analysis, so the same marital status variable as in approach 2 was used. There were 171 participants who had a report of being unmarried with a child before the age of 18 and before 2014. As the number was relatively small, cluster analysis was not carried out and the sequences were simply displayed on index plots by calendar era of the start of the sequence (80 in 2014-2008 and 91 in 2009-2013), there was an attempt to order the sequences by the number in each state to aid interpretation: the plots are shown in figure 7.5 and descriptive statistics in table 7.6. Initial observations show that the earlier era seems to be affected by a greater proportion of participants leaving the area for other reasons (40% vs. 30%) and the later era has more participants who experience divorce (17.6% vs. 10%).

As well as the findings reported above, the sequence analysis in this approach provided a lot of useful information at a glance which would help an investigator design a study, for example that while some participants remain unmarried, most do marry at some point. It also helps to highlight potential data issues: for example, the few participants marrying in the next quarter after the birth, or one experiencing divorce without having been married may need some investigation.

Figure 7.5. Sequence index plots for 5-year sequence of marital status following having a child without being married by calendar era

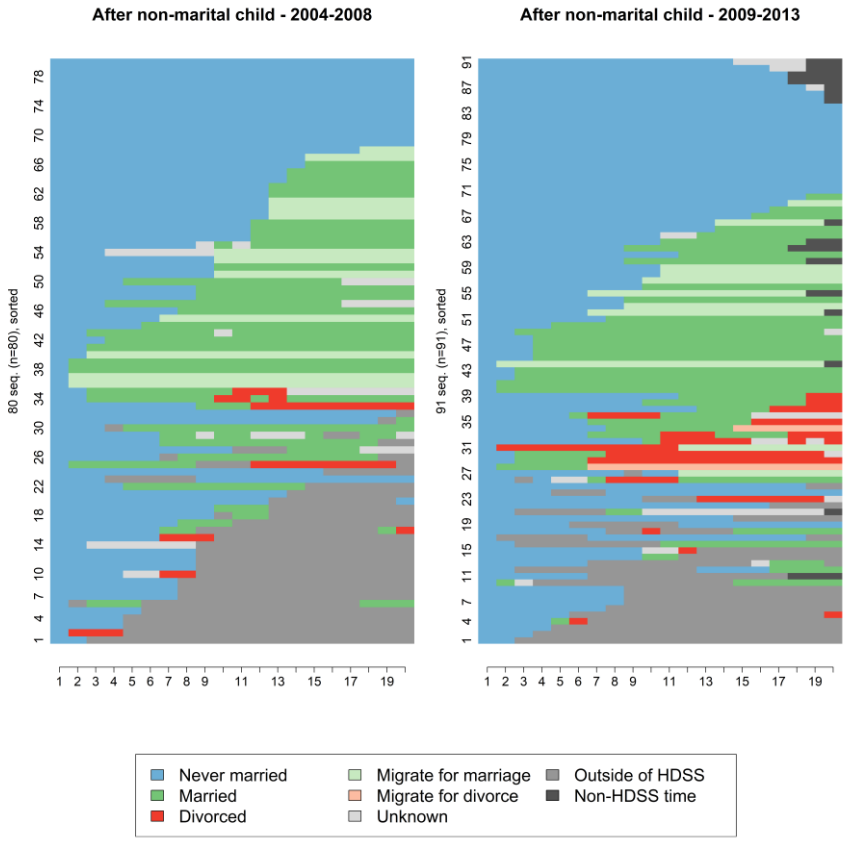


Table 7.6. Descriptive statistics for 5-year sequences in marital status following first report of a birth while unmarried and aged under 18 years by calendar era. Overall number (and proportion), and mean (and standard deviation) per person, of quarters spent in each category is shown.

	2004-2008 (n=80)		2009-2013 (n=91)	
	Num (%)	Mean (sd)	Num (%)	Mean (sd)
Never married	80 (100.0%)	9.3 (6.3)	91 (100.0%)	10.1 (6.5)
Married	37 (46.3%)	4.5 (6.0)	42 (46.2%)	4.2 (5.9)
Divorced	8 (10.0%)	0.4 (1.5)	16 (17.6%)	0.9 (2.6)
Migrate for marriage	12 (15.0%)	1.7 (4.5)	11 (12.1%)	1.2 (3.6)
Migrate for divorce	0 (0.0%)	0.0 (0.0)	2 (2.2%)	0.2 (1.6)
Unknown	10 (12.5%)	0.5 (1.5)	14 (15.4%)	0.4 (1.3)
Outside of HDSS	32 (40.0%)	3.6 (5.5)	27 (29.7%)	2.6 (4.8)
Non-HDSS	0 (0.0%)	0.0 (0.0)	16 (17.6%)	0.3 (0.8)

7.5. Discussion and conclusions

This investigation into the use of sequence analysis to study life course transitions using HDSS data has demonstrated both strengths and weaknesses. Although caution is required and a careful understanding of potentially biases due to migration of participants, it can be used to draw conclusions regarding the transition to adulthood and whether it has changed over time. The results suggest that for many women in the study the transition to adulthood experienced is quite traditional: leaving home to marry and rapidly have children; but that there is some evidence that there has been some change over time, with marrying early becoming less likely and remaining in school and staying in school for longer becoming more common. Sequence analysis can certainly be used as an exploratory tool to generate ideas for research questions, and in particular to aid in visualisation of available data: for HDSS dataset this is useful as movement in and out of the area is so common.

7.5.1. Utility of sequence analysis

The selection of the sequence and cluster analysis methods and algorithms can be confusing for the novice sequence analyst. I spent quite some time trying to understand the different issues and was fortunate to be able to use the 2016 review of dissimilarity measures to help decide which to use (Studer and Ritschard, 2016). The fact that there were 18 different methods to review in this paper demonstrates the complexity of the issue and the difficulties that non-experts have in discerning the methods to use. However, that paper's recommendations for the methods to use with multichannel sequence analysis were actually not available within the TramineR R package, so I opted to use the simplest Hamming distance method, as this seemed appropriate to examine the timing of transitions and also allowed for user-defined substitution costs for the single-channel approach with the migration states. I was keen to use these so that the specific migration states would be more likely to cluster with the known equivalent marital status state. The assigned costs are arbitrarily assigned however, so are sometimes discouraged. Similarly, there has been much debate and criticism over the method used to create the clusters. Wards algorithm is often used, though criticised and other methods suggested (Lesnard, 2006), however I used Wards as the other techniques produced confusing and unhelpful results. Hierarchical clustering methods as a whole have been criticised as the clusters do not rejoin once split (Liao et al., 2022), I did not attempt to try any other data-driven methods of data portioning, but this might be something to look at in the future. I did manually re-join clusters when it seemed appropriate as this has been done before: a multi-channel sequence analysis looking at employment, housing, marital and family status created 15 clusters and, rather than rely on

the cluster quality statistics, manually re-joined some of the splits following examination of the clusters (Pollock, 2007). I also use manual techniques to create clusters and showed the utility of visualising sequences by other existing groups, which has also been shown to be useful and valid before (Liao et al., 2022).

I found the multi-channel sequence analysis useful, however the number of graphs to examine and display can become unmanageable, especially if the number of clusters are high: I deliberately chose the lowest reasonable number of clusters despite the cluster statistics suggesting a higher number, one of the reasons for this was that interpreting 8 groups across 4 domains would be challenging. It has been reported that multi-channel sequencing can sometimes be dominated by one particular domain, especially if that domain is more turbulent than other (Liao et al., 2022), as my domains tended to following similar timing patterns this was not something I had to deal with.

Dealing with missing data is a current priority for sequence analysis researchers. One example experimented with assigning the cost of missing during the creation of the dissimilarity matrix, but did not come to a totally satisfactory conclusion. They did find that including missing as a standard state tended to create clusters based on missingness, which is what I experienced. They suggested that the analysis may need to be carried out separately according to the length of sequence available (Lazar et al., 2017). Imputation methods for missing states exist, however are mostly useful for short periods of missingness in the middle of sequences (Liao et al., 2022). For my analysis, the missingness is likely related to the outcomes so it would probably not be appropriate to attempt imputation. I am not aware of another analysis which assigned different categories of missingness as I did.

I used sequence index plots throughout this report, these are figures which display each individual sequence as one line with different colours representing the states. These are most useful for up to about 400 sequences per plot as for bigger groups the lines become too thin to distinguish: for larger groups it is recommended to display a random sample or just to display the most common sequences (though the latter will obviously make the group appear more homogenous than it is) (Brzinsky-Fay, 2014). Most of the clusters that I generated had under 400 members and I did not need to see the detail in the larger groups i.e. the group that remained unmarried most of the time had some changes in status right near the end of the period which were indistinguishable on the plot, however for my purposes it would not change the interpretation, so I did not attempt to view a selected sample. Other ways of visually displaying the information are available, including graphs which show the proportion of in each state at each time point (which do not have the issue of

groups with different number of members displaying differently) and modal plots which display the most common category at each time point (Brzinsky-Fay, 2014) I find the former to be helpful, though the index plots are vital to fully understand the dynamics of the group. I did not attempt to use modal graphs as these would not be helpful for the interpretation of the clusters, however they may be more useful for display in outputs.

For this preliminary assessment of using sequence analysis on HDSS data I focussed on the visualisation of the sequences and basic summary measures. There are also other summary measures that can be generated to look at the diversity and complexity of sequences, and through user-defined inputs, sequences can be assigned a summary measure of favourableness (or unfavourableness) (Ritschard, 2021). For my analysis, I felt the visualisation and basic descriptions of each cluster with adequate as there did not seem to be much evidence of turbulence or too much complexities within the sequences I was looking at, however some of these approaches might be useful for further investigation of the participants, for example, who experience divorce. Additionally, a recent paper suggested a dynamic method of assessing complexity of sequences which reduces the need to focus so strictly on a specific birth cohort, allowing people with only partial data available to contribute (Pelletier et al., 2020): this may have useful applications for HDSS data. Current developments looking at how to include other time varying covariates as well (Liao et al., 2022). Another potential aspect of sequence analysis that could be useful for HDSS data is looking at linked sequences: this enables comparison of sequences between dyads, for example a recent study demonstrated how sequences of family formation can be compared between parents and children (Liao, 2021). The Karonga HDSS has excellent data on linkages between families so this could be a useful avenue to explore.

7.5.2. Transition to adulthood

My results here have shown that for many women in this rural location in Malawi, the transition to adulthood follows a quite traditional trajectory: leaving school and home to get married, and then rapidly building a family. There are few comparable studies in sub-Saharan Africa: one analysis in urban South Africa, found a lot of heterogeneity in pathways taken (Biddecom and Bakilana, 2003) while another in a similar urban location was able to identify standard trajectories (Bennett and Waterhouse, 2018), though they were quite different from those found in the present study, which is not surprising as urban inhabitants have different opportunities. Our analysis also identified a sizable group of women are not marrying before their early 20s and are remaining in education, and there is evidence that this group has increased in size over time. These findings reflect what has been observed elsewhere: since the year 2000 there has been a large global increase in secondary and

tertiary education attendance and an increase in the average age at marriage and age at having a child for women (Juárez and Gayet, 2014). The decrease in groups marrying early and increase in those marrying later also confirms what was found in the large study on transition to adulthood using sequence analysis on data from 69 countries: in the East Africa group they found that their 'rapid early transitions' category decreased and 'delayed rapid transition' increased over birth cohorts (Pesando et al., 2021). All 3 analyses showed that a proportion of the young women experienced divorce, often not long after first getting married, and a good number marry again. This confirms previous findings in Malawi which showed that divorce was common in the first three years of first marriages (Bertrand-Dansereau and Clark, 2016), and that in general remarriage happens very quickly after divorce (Malinga John, 2022).

My analysis did not include the transition to work as the area is rural and many people may report being farmers even while still in education, so do not experience a clear transition from education to work. It has been a concern that the global increase in access to education has not uniformly improved access to secure and adequately paid work and that a high proportion of young people experience periods of unemployment (Juárez and Gayet, 2014). A study in nearby Zambia found that many young people are spending a long time establishing themselves in the world of work (Locke and Lintelo, 2012). Most of the women in my analysis who are not married are in school so there was no evidence that they are experiencing an unwanted delayed transition to adulthood or unemployment. I did not include men in this analysis, as the focus was more on assessing the sequence analysis technique, but including them in future might shed light on whether there is evidence of unwanted delays in transition to adulthood in this area. The women leaving the area without being married in our analysis may be leaving for education, work, or simply moving with their families. Lack of information on this group does hinder making conclusions about the transition to adulthood in this area as this could be a group of people leaving home to live independently in urban areas without being married, which would change the conclusion about the traditional transition to adulthood: leaving home has not be studied as much as other markers of adulthood, but in some areas the age of leaving home has decreased, possibly due to increasing need for young people to migrate to urban areas to find work (Juárez and Gayet, 2014).

7.5.3. Strengths and limitations

While this report highlights some of the issues with using HDSS data to examine individual longitudinal trajectories, particularly when the subject under investigation is likely related to leaving and entering the study area, it also has some benefits. Firstly the flexibility of data

manipulation within HDSS dataset allows for different lengths of period to be generated depending on the research question: in my analysis I used quarterly snapshots, but monthly, 6-monthly or yearly snapshots could also be created. HDSS event dates should be quite accurate as they are collected prospectively so sequence states created from them (i.e. timing of births, deaths, house moves and household membership) should allow for quite specific fine-grained analysis. Other information collected through regular surveys (i.e. school attendance or marital status) may not have such accurate dates attached to them so using shorter sequence periods would not add anything. The difference in accuracies of the dates used to construct the sequences in my analysis is one of the limitations of this analysis: the marital status variable was constructed from annual reports of marital status and dates of marriage for specific spouses. In some cases this lead to inconsistent data, i.e. a period of 'never married' following one of 'married'. In such cases the inconsistent data was over-written with the most likely state (i.e. 1 report of 'married' within a period of 6 'never married' reports would be over-written as 'never married', while a reported of 'never married' following a period of 5 'married' reports and preceding 2 reports of 'divorced' would be over-written with 'divorced'). This data cleaning may have lead to some inaccuracies, for example as highlighted in the third analysis where it could be seen that some participants transitioned straight from never married to divorced.

As described in the results section, using different categories for missing data may create an artificially high feeling of confidence about the results. In this particular case the uncertainty over whether the reason for moving applied to the participant or someone else also added to this limitation. However, equally, it does make sense to utilise all available data, and I felt it had some use.

Another benefit of using HDSS data for this type of analysis is the scope to link to other past or future data outside of the sequence period under analysis: as it has been shown previously that frequent changes in family living arrangements during childhood is associated with following pathways to adulthood which may be considered less ideal (Goldberg, 2013) and that different trajectories to adulthood are associated with future health (Bennett and Waterhouse, 2018). This sort of analysis however would also be subject to the same biases related to migration as demonstrated in this report so careful selection of time periods and examination of people included would be required.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	145184	**Title**	Ms
First Name(s)	Estelle		
Surname/Family Name	McLean		
Thesis Title	Demonstrating the value of Health and Demographic Surveillance Site data for complex secondary analyses, illustrated with analyses of young people's living arrangements and transitions to adulthood.		
Primary Supervisor	Rebecca Sear		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Wellcome Open Research
Please list the paper's authors in the intended authorship order:	Estelle McLean, Albert Dube, Fredrick Kalobekamo, Emma Slaymaker, Amelia C Crampin, Rebecca Sear
Stage of publication	**Submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I developed the concept for the analysis, conducted all the data manipulations and analyses, wrote the manuscript and revised according to comments from co-authors

SECTION E

Student Signature	
Date	27 July 2023

Supervisor Signature	
Date	27 July 2023

8. Local and long-distance migration among young people in rural Malawi: importance of age, sex and family

Estelle McLean*[1,2], Albert Dube[1], Fredrick Kalobekamo[1], Emma Slaymaker[2], Amelia C Crampin[1,2], Rebecca Sear[2]
*author for correspondence: estelle.mclean@lshtm.ac.uk
1. Malawi Epidemiology and Intervention Research Unit, PO Box 46, Chilumba, Karonga District, Malawi
2. London School of Hygiene and Tropical Medicine, Department of Epidemiology and Population Health, Keppel Street, London, WC1E 7HT

8.1. Abstract

Background In sub-Saharan Africa, migration of young people is common and occurs for a variety of reasons. Research focus is often on international or long-distance internal migration; however, shorter moves also affect people's lives and can reveal important information about cultures and societies. In rural sub-Saharan Africa, migration may be influenced by cultural norms and family considerations: these may be changing due to demographic shifts, urbanisation, and increased media access.

Methods We used longitudinal data from a Health and Demographic Surveillance Site in rural northern Malawi to present a detailed description of short and long distance, independent and accompanied, migration in young people between 2004-2017. We further explore the family and household factors which are associated with these moves using multi-level multinomial logistic regression modelling.

Results & conclusions We found two key periods of mobility: in very young childhood, and in adolescence/young adulthood. In this traditionally patrilocal area, we found that young women move longer distances to live with their spouse. Despite the local patrilineal customs, we found evidence of the importance of the maternal family, and that female and male children may be treated differently from as young as age four, with girls more likely to migrate long distances independently, and more likely to accompany their mothers in other moves. Young people living close to relatives tend to have lower chances of moving, and those from more advantaged families are more likely to move.

8.2. Introduction

Research in sub-Saharan Africa has found relatively high rates of mobility/migration in young children and adolescents/young adults (Beegle and Poulin, 2013; Ford and Hosegood, 2005; Grieger et al., 2013). Children and adolescents may move with their parents/guardians, and adolescents/young adults can be expected to move away from their natal home to live independently. Dependent children or adolescents may also move without their parents or guardians: children in sub-Saharan Africa may be fostered out to other households either temporarily or long-term, so that the fostering household may provide care for them or give them better educational opportunities, or so they can provide support to the fostering household through house/farm work (Hedges et al., 2019) or caring responsibilities (Robson, 2000). Fostering may happen in a planned way, or in more emergency circumstances due to parental illness or death. Older children/adolescents may be sent to live in other households where there might be better able to work and earn wages to send back home (Kwankye, 2012; Temin et al., 2013).

On a population level, it is important to study migration to understand population flows and predict needs for services. It is also important to understand the effects of migration on individuals and communities as both positive and negative outcomes have previously been found: Migration can give people opportunities for education, work and to form new relationships: for adolescents in some communities, a period of migration has become a common part of transition to adulthood, where they can learn new skills and enjoy their independence (Hertrich and Lesclingand, 2013); However, migration, particularly in unaccompanied children and adolescents, is often regarded in a negative light, as migrants may be more likely to drop out of school (Clark and Cotton, 2013; Hashim, 2007), experience increased work load (Hashim, 2007), social isolation (Temin et al., 2013), pre-marital pregnancy (Xu et al., 2013) and HIV (Anglewicz and Reniers, 2014) (though it has been suggested that vulnerabilities associated with migration also lead to increased risk of HIV (Magadi, 2013)); in Uganda, for both young men and women, migration was associated with greater chance of recent alcohol use and increased likelihood of engaging in risky sexual behaviour (Schuyler et al., 2017).

Migration is often studied in respect to international movements, and if internal movements are considered then often only long-distance moves are counted. Moving locally compared to long distance may have fewer consequences for people and the local environment: a qualitative study in Malawi found that when asked about moving house, children did not tend

to report on short moves when they stayed with family or moved to a different family member, and the authors speculated that this could be because these moves did not change their lives much (Young et al., 2006). However, a local move is still a potentially disruptive event which may have both positive and negative effects (van Blerk and Ansell, 2006) on a young person's life, and short-distance migrations can also give valuable insights into local cultures and customs. In Malawi, international migration, in particular to South Africa, was common in colonial times but was discouraged from the 1960s (Beegle and Poulin, 2013). While in recent years international migration has increased once more, internal migration is far more common (Beegle and Poulin, 2013), and a high level of circular migration has been found among adolescents (Chalasani et al., 2013).

8.2.1. Theoretical background and literature review

Theories of migration have been developed over the past several decades which attempt to explain migration patterns through economic, demographic, human capital and risk diversification lenses (see Piguet 2018 for a summary). Many general concepts can be applied to the migration of children and adolescents in sub-Saharan Africa, in particular the New Economics of Labour Migration theory, which places the household rather than the individual as the decision-making unit (Stark and Bloom, 1985). There are a few specific points to note regarding this household-based framework when conceptualising migration of young people in rural sub-Saharan Africa. Firstly, there is the level of agency that children have over household decisions that affect them. Of course, decision-making power may not be equally spread among household members of any age, but this is particularly the case among children, whose agency within the household depends on multiple factors including age and sex (boys may have more say in migration decisions than girls (Hunleth et al., 2015)). While the adults in the household are likely to make the final decision over the migration of a child or young adolescent, it would be incorrect to assume that they have no agency in the decision-making process: for example, after a divorce children may decide themselves who to accompany or may attempt to negotiate staying with relatives in the neighbourhood even if their parents move. Equally those children 'sent' away may be given different levels of say over this decision, even if the initial idea was not theirs (Hashim, 2007). The second factor is related to adolescents' desire for independence, and how the timing or format of this may be at odds with the desires of the other household members. Whitehead and colleagues suggest that migration in children and adolescents, particularly independent migration, can be viewed through the lens of the intergenerational contract: shared understandings between family members of what may be expected from each other (Whitehead et al., 2005). Thirdly, defining a household is not straightforward and is subject to cultural influences (Randall et al., 2011); this has been shown to affect understanding of

migration in Nepal (Agergaard, 1999). Finally, extended family is important in many cultures and it is likely that migration decisions will be made taking into account not just the immediate household or family, but other family members. Research in Zambia suggest that, rather than migration being the result of the household deciding what is best for all members, it can often be the result of conflicts over land and resources, leading to rupture in the household (Cliggett, 2000).

Research on mobility and migration in young people in Africa tends to find differences by sex. Female children have been found to be more likely to be fostered out in research in Tanzania (Hedges *et al.* 2019) and Zimbabwe (Robson 2000). In Zambia, a study looking at short-term movements, found that girls were more likely to spend the holidays with extended kin, and also more likely than boys to have the reason for going as 'helping with household chores'; boys were more likely to report 'getting to know relatives' though this was by far the most common response for both sexes (Hunleth et al., 2015). In general, adolescent girls and young women tend to be more likely to move: in Kenya, young migrants to urban areas were more likely to be female (Clark and Cotton, 2013), in Malawi a study of young people aged 15-24 found that 47% of women had ever moved compared to 38% of men (Beegle and Poulin, 2013). There are also differences found in the reason given for the move: young women tend to report moving for marriage, and men for work, education or economic reasons (Anglewicz and Reniers, 2014; Beegle and Poulin, 2013; Chalasani et al., 2013; Clark and Cotton, 2013) however it has been suggested that sex differences in reporting reasons for move may simply reflect the gender norms of the society, *i.e.,* it may not be socially acceptable for a woman to report moving for economic reasons, so she reports marriage even if her main consideration was economic, and vice versa for men (Temin et al., 2013). Despite these reporting biases, economic migration in young women has been seen to be just as common as in men in South Africa (Camlin et al., 2014) and in many West African countries a period of migration has been part of the transition to adulthood with young women usually spending time in the city before returning for marriage. This phenomena tends to be viewed differently for boys than girls, with girls facing more reluctance from their elders to let them go, the authors speculate that migration in young males is in line with family expectations (*i.e.* to provide) so is accepted, while for women it is perceived that it mostly benefits the individual rather than the family (Lesclingand and Hertrich, 2017), despite many young women report using the time to learn domestic skills and earn money to buy items to help them in their marital home (Hertrich and Lesclingand, 2013).

Socioeconomic position has also been found to be associated with moving, though it seems to have a complex relationship with migration, with evidence that the most and least disadvantaged groups are most likely to move (Ginsburg et al., 2009). In Malawi, those in wealthier households were more likely to move unless they were farming families (Beegle and Poulin, 2013). In Senegal, young people whose fathers had more education or higher socio-economic position in childhood were more likely move to urban areas, but less likely to move to rural ones: however this was influenced by the initial location, with the author reporting that access to more and better community resources makes later moves less likely (Herrera-Almanza and Sahn, 2020). In Malawi, young movers were more likely to come from households with more assets (Chalasani et al., 2013).

Several studies have also examined parental presence, vital and marital status and household composition and found these to be associated with youth migration: children were found to be more likely to move if not living with their mother (Madhavan et al., 2012) and adolescents were more likely to move if the household head was not a parent (Beegle and Poulin, 2013; Chalasani et al., 2013). Loss of parent has also been found to increase migration for young women in Senegal (Herrera-Almanza and Sahn, 2020), but in Malawi this effect was only seen in men (Beegle and Poulin, 2013). While most studies have focused on parents, some have attempted to look beyond this *i.e.* Clark *et al*. in Kenya asked young respondents to indicate who was responsible for them, rather than making assumptions (Clark and Cotton, 2013) and a Malawian study asked generally about family and friends in the area. This latter study found that knowing family and friends prior to moving was associated with a longer length of stay at that location, but especially for women, knowing friends there was the most strongly associated (Myroniuk, 2018). In many sub-Saharan African societies, the extended family is important in day-to-day life and is likely to impact decision making around migration.

8.2.2 Aims of this paper

This analysis aims to provide a detailed description of mobility and migration in children and adolescents/young adults in rural Malawi, and to assess the role of family (within and outside the household) on accompanied and independent, long-distance and local migration in the same population using multinomial multi-level regression modelling. The data is from the Karonga Health and Demographic Surveillance Site from 2004-2017 and allows examination of the presence of different types of family members both in the household and living nearby: we have previously shown that young people in this population tend to live near extended family (McLean et al., 2021a).

8.3. Methods

8.3.1. Context

The Karonga Health and Demographic Surveillance Site (HDSS) was established in 2002 in the southern part of the Karonga district in northern Malawi (Crampin et al., 2012) by the Malawi Epidemiology and Intervention Research Unit (MEIRU- formerly known as the Karonga Prevention Study). The area is largely rural with one semi-urban trading town, several smaller market villages and one port on Lake Malawi. The majority of the population engage in subsistence farming or fishing. The main ethnic group are Tumbuka, who have followed patrilineal and patrilocal custom since the 19th century: women tend to move to their husband's village when they marry (Malawi Human Rights Commission 2006). In the event of divorce or even paternal death, children considered to be old enough to be away from their mother may be required to live with their father's family (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.

The HDSS covers an area of 150km^2 and by 2016 had over 40,000 people under surveillance, with very high response rates. Household membership is defined by the participants with guidance from trained fieldworkers: all household members must usually live in the dwelling/compound together and recognise the same household head. Men with more than one wife who do not live in the same location are assigned to be living in each wife's household; all other individuals may only belong to one household. Households are identified by two unique numbers: one which does not change through the lifetime of the household (known as unique household ID for this analysis) and one which is related to the household's location, which may change over time (known as geographic household ID for this analysis); GPS coordinates are recorded for each household when they are registered and if they move.

Births and deaths are captured monthly through a system of local 'key informants', while migrations are captured annually through visits to all households: information is gathered on any new or departed household member including date of move, reason for move and where they moved from/to. If a whole household moves, then this information is gathered from the key informants. When a new household member 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 (even if they are not currently HDSS participants). 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. This information was used to identify all family links (by blood and by marriage) between all HDSS participants.

8.3.2. Ethics

Household heads provide written informed consent on behalf of the whole household to participate in the Karonga HDSS, which may be rescinded at any time for any reason. The HDSS is regularly reviewed and approved by the Malawian National Health Sciences Review Committee (approval #419), and the London School of Hygiene and Tropical Medicine Ethics Committee (approval #5081).

8.3.3. Dataset

Data on HDSS participants are gathered as event reports and surveys. The event data is used to create continuous episodes, and the survey data are assumed to be valid for dates within certain periods before and/or after the survey date (the length of these periods depend on the type of data). Generation some of the exposure variables involves calculating the distance between each index and multiple relatives and would be computationally intensive to apply to the longitudinal dataset so the episodic data were reduced to one data point per quarter (15th of the middle month of each quarter) per person, and the data treated as panel data. Each data point included variables indicating whether a person was living with, or within 250 metres of, specific types of family member, various indicators of socio-economic status, and information about the local area.

Moves were identified as individuals who had a different geographic household ID to the following quarterly snapshot, or if they were not present in the HDSS in the next quarter and were recorded to have migrated out, or if they were not present the previous quarter and were recorded as migrating in. Moves of less than five metres were dropped as these were likely to be artefacts of a new geographic household ID being assigned if a new household head was declared. If there was more than one move associated with a quarter (*i.e.* a move in from outside of the area, and then a move out) then one move was kept randomly. Due to some disruptions to data collection in recent years, including due to Covid-19, only data up to the end of 2017 was used for this analysis.

8.3.4. Exploratory analyses

The analysis can be divided into three main steps: exploratory, descriptive and regression analyses. *A priori*, it was decided to analyse the moves (1) by distance, (2) whether they

were independent or accompanied, and by (3) age and (4) sex. The cut-offs/definitions used were defined during the exploratory analysis as described below.

Distance

When a person moved within the HDSS area the actual distance moved was calculated using household coordinates, which are available for all. When the move was to or from an area external to the HDSS, in almost all cases the source or destination was gathered: for a town or city in Malawi, GPS coordinates from a central area of the town/city were used, for outside of Malawi, GPS coordinates of a point in the new country nearest to Malawi were used. This means that our analysis includes both internal and international migration, though the majority of moves are internal. To define the cut-off between short and long moves, a preliminary analysis of primary school attended before and after a move was carried out. Data are captured annually to record school grade and school name for school attenders: records of all primary school attenders were examined if they were still attending primary school and had a different geographic household ID (*i.e.* had moved house) at the following interview. The distance between the households of the first and second interview was calculated and the averages compared between those who changed school and those who remained at the same school. There were 3916 record pairs from 3010 individuals which met these criteria. 2490 did not involve a change in school and the mean distance moved was 0.9km (95% CI allowing for clustering by individual ID 0.86-1.03), 1426 did change school and the mean distance was 4.6 (95% CI allowing for clustering by individual ID 4.4-4.8). Based on these results, it was decided to categorise short move as less than four kilometres and long moves as four kilometres or more.

Independence

To identify whether the person moved alone or with members of their household, all household members were assessed to see whether they stayed in the original house, moved with the index person, or moved elsewhere. For external moves they were classed as moving together if they reported the same source or destination town or country, for moves within the HDSS they had to have the same destination household ID. People were then classed as moving independently (without a parent of any age, or an adult aged 18 or over) or accompanied (with at least one adult aged 18 or over, or a parent who may be under 18).

Age/sex

The above definitions were applied to the panel dataset so that each record had an outcome of either ‘no move’ or one of four move types (short independent, long independent, short accompanied or long accompanied). The risk of each move outcome was calculated for

each age year, separated by sex. Although the focus of the analysis was on young people, a high upper age limit of 34 was used initially, to be sure of observing the whole of the adolescent/young adult time. 95% confidence limits were calculated allowing for clustering within unique household ID and unique individual ID. Following these initial analyses (which are described in the results section), age was categorised as ‘children’ for females if under the age of 12 and for males if under 16 and ‘adolescents’ for females if aged 12-24 and males 16-28. The age ranges are different as females tend to experience transitions to adulthood earlier than males in this area (McLean et al., 2021b); even though the age ranges extend beyond typical definitions of adolescence, the term adolescent will be used for simplicity throughout.

8.3.5. Descriptive comparison of family/household composition of movers by sex

For this analysis, the panel dataset was modified to only include records with a move outcome. Whether the household composition (the make-up of the household members in terms of their relationship to the index mover) differed between the sending and receiving household was examined using only short moves, as these are most likely to have full information on both sending and receiving household (as most long moves were to or from outside of the area). Firstly, the numbers moving between different household types were displayed on Sankey diagrams (created using the networkD3 package in R (Gandrud et al., 2017)) separately for children and adolescents, male and female, and independent and accompanied moves. The proportions in each category of the sending and receiving households were then compared by sex, separately by move type and age group, using Wald tests allowing for clustering within unique household ID and unique individual ID. The household composition variable was created using latent class analysis which has been described elsewhere (McLean et al., 2021a) with the following categories:

- Parents & siblings – both parents present, plus siblings aged under 18 if any, and does not fit into any of the below categories
- Sister's family – at least one sister aged over 18 years or her family, (more of sister+family present than brother+family if also present), and mother and/or father present or no maternal or paternal family present
- Brother's family – as above but with brother instead of sister
- Mother & siblings – mother present, no father present, nor father's other wife nor maternal family
- Father & stepmother – mother not present, father or father's other wife present
- Maternal – father and father's other wife not present, at least one maternal relative present and more maternal relatives present than paternal

- Paternal – mother and father's other wife not present, at least one paternal relative present and more paternal relatives present than maternal
- Spouse – spouse present
- Other – does not fit into any of the above categories
- External – household composition unknown as outside of HDSS area
- No IDs – household composition unknown as parent IDs unknown

In addition, whether children and adolescents were moving with their parents was examined using all short and long accompanied moves: the proportions in each category (no parents, mother only, father only, both) were compared by sex, separately by move type and age group using Wald tests allowing for clustering within unique household ID and unique individual ID.

8.3.6. Regression analysis of associations between mobility and family and household composition/structure

For this analysis, the full panel dataset was used, including those who didn't move. However, moves with no information on the 'sending household', *i.e.,* those migrating from outside the HDSS, were dropped, and, as the focus was on the effect of family, participants who had no record of either parent ID were excluded. Multi-level multinomial logit regression models allowing for clustering by unique household ID and unique individual ID were run on the full panel dataset using MLwiN Version 3.05 (Charlton et al., 2020), with the outcome of move type (no move [baseline], short independent, long independent, short accompanied and long accompanied) and including the family variables and potential confounders. The model was run separately for each age group and sex. The potential confounders were chosen due to existing literature having demonstrated associations with mobility (see introduction), and data availability (*i.e.* further household socio-economic status variables could not be included due to lack of complete data). The variables included in this analysis were:

- Time-varying variables relating to family and household composition/structure:
    - A detailed household composition variable (as described above).
    - Number of people in the household in different age groups (under one year, one-four years, five-11 years, 12-18 years, 19-29 years, 30-59 years and 60 years and over) (all continuous variables which exclude the index person)
    - A total of four binary variables indicating presence of at least one relative within 250 metres but not in the immediate household from five family types: maternal (not including mother), paternal (not including father), sister's (including sister aged 18 or over, but not younger sister), brother's (including

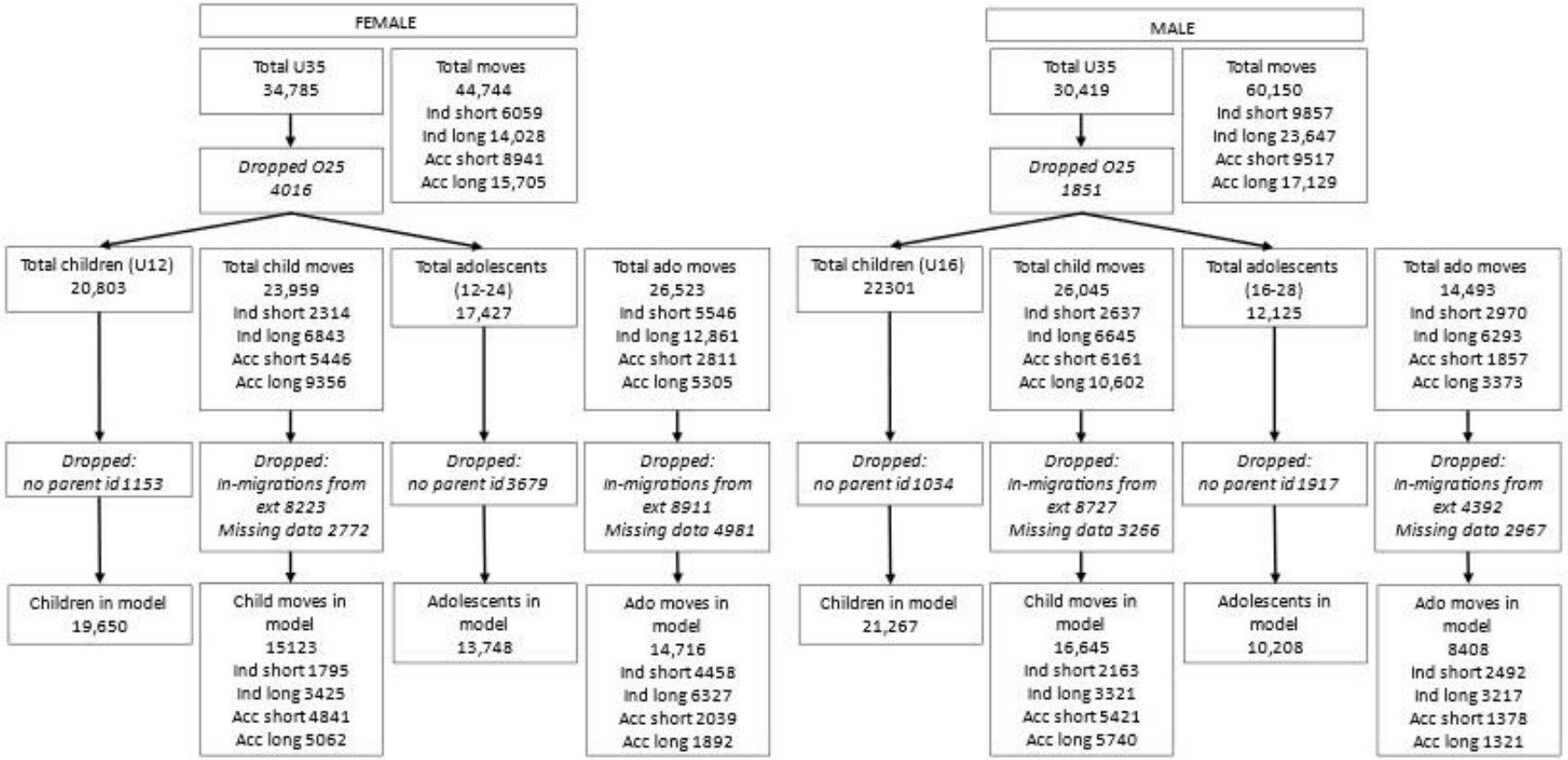
brother aged 18 or over, but not younger brother) and nuclear (parents and siblings aged under 18). The distance of 250 metres was chosen as it was assumed that relatives living closely are likely to be seeing each other regularly and have some influence on each other's lives.

- Potential confounding variables:
    - Presence of own child in the household: to allow assessment of presence of small children after accounting for own child as having an own child may affect migration decisions (only for the adolescent analyses)
    - Whether biological mother is known to be dead (vital status of parents is derived directly from the HDSS record of the parent, or from information gathered on each participant's parents at the annual survey)
    - Whether biological father is known to be dead
    - Age
    - Year (in two-year bands)
    - Distance to tarmac road,
    - Population density within 250 metres (categorised),
    - Whether father had any secondary education,
    - Whether mother had any secondary education
    - Employment ranking of the household head (categorised into low, which include not working or precariously employed, medium which includes subsistence farming and fishing, and high which includes waged employment and business owning)

8.4 Results

A total of 65,204 (34,785 female and 30,419 male) individuals aged under 35 contributed data to the initial exploratory analysis. There were 104,883 (44,733 female and 60,150 male) moves: 15,916 short independent; 37,675 long independent; 18,458 short accompanied and 32,834 long accompanied (Figure 8.1).

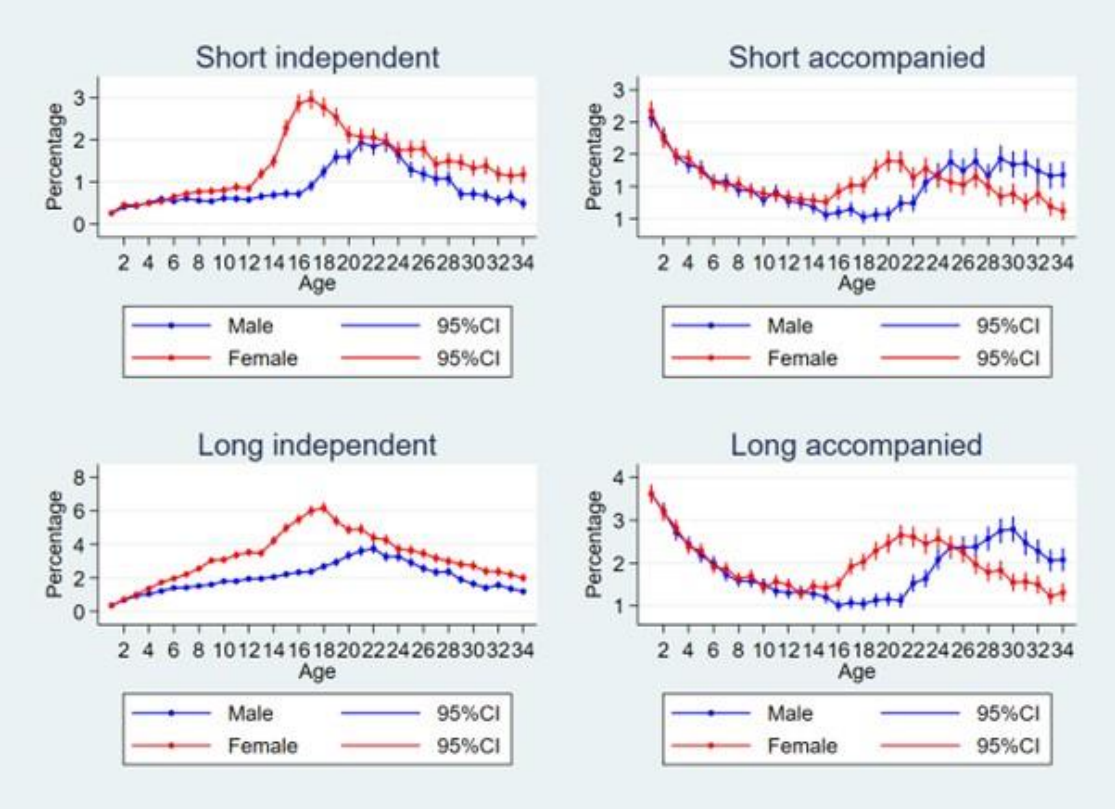
Figure 8.1. Number of participants and moves at different stages of analysis



8.4.1. Percentage who moves by age and sex

Plots of the percentage of individuals that experienced each move type (short independent, long independent, short accompanied and long accompanied) at difference ages, by sex can be seen in Figure 8.2. For short independent moves, the risk increased sharply for females from the age of 12 up to age 17 before falling less steeply; for males the risks only increased from age 16, peaking at age 21-23, however risks were almost always lower than for females. For long independent moves for both sexes, the risks gradually increased during childhood, for females the risk was higher than for males from the age of four onwards, and the increase became steeper age 12 with a peak at age 18; for males the peak was at age 22. For short, accompanied moves, the risks were highest in the youngest children (risks for females and males are similar at very young age) which then declined to age 14-15, from here the risks increased again for females to a peak at age 20, while for males the risks increased again to age 24 and then remained stable or declined. The pattern was similar for long accompanied moves, except that for males the risks continued to increase up to age 30, before declining.

Figure 8.2. Percentage of individuals experiencing each move type at difference ages, by sex



NB: Short moves are under 4 kilometres; independent moves are without a parent of any age or an adult aged 18 or over; male data is blue, female is read; note that each figure has a different scale

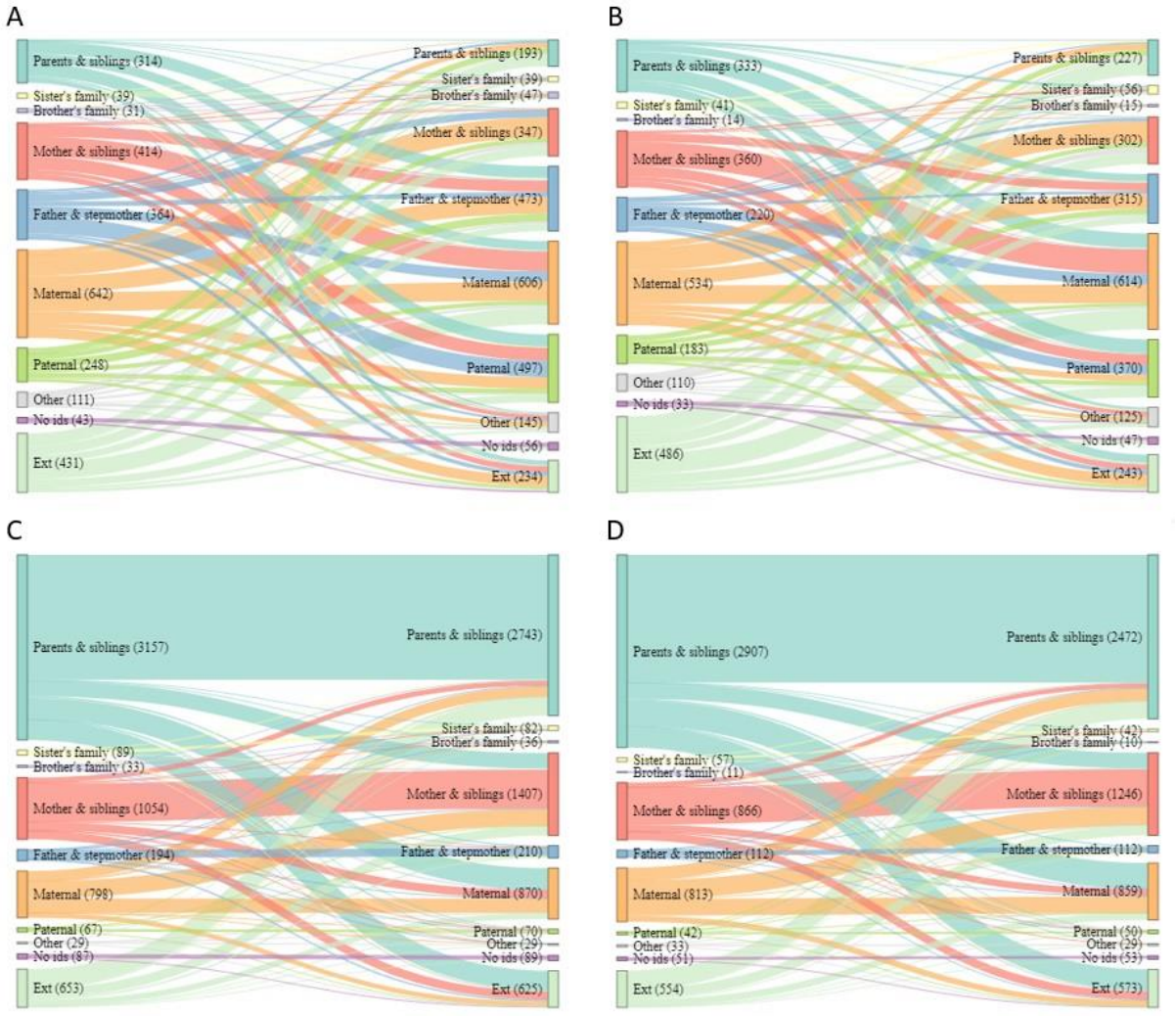
8.4.2. Comparison of family/household composition of movers by sex

A total of 43,104 (20,803 females aged under 12, and 22,301 males aged under 16) children and 29,552 (17,427 females aged 12-24 and 12,125 males aged 16-28) adolescents contributed data to the descriptive analyses. There were 50,004 (23,959 female and 26,045 male) child moves (4951 short independent; 13,488 long independent; 11,607 short accompanied and 19,958 long accompanied) and 41,016 (26,523 female and 14,493 male) adolescent moves (8516 short independent; 19,154 long independent; 4668 short accompanied and 8678 long accompanied) (Figure 8.1).

8.4.3. Changing household composition

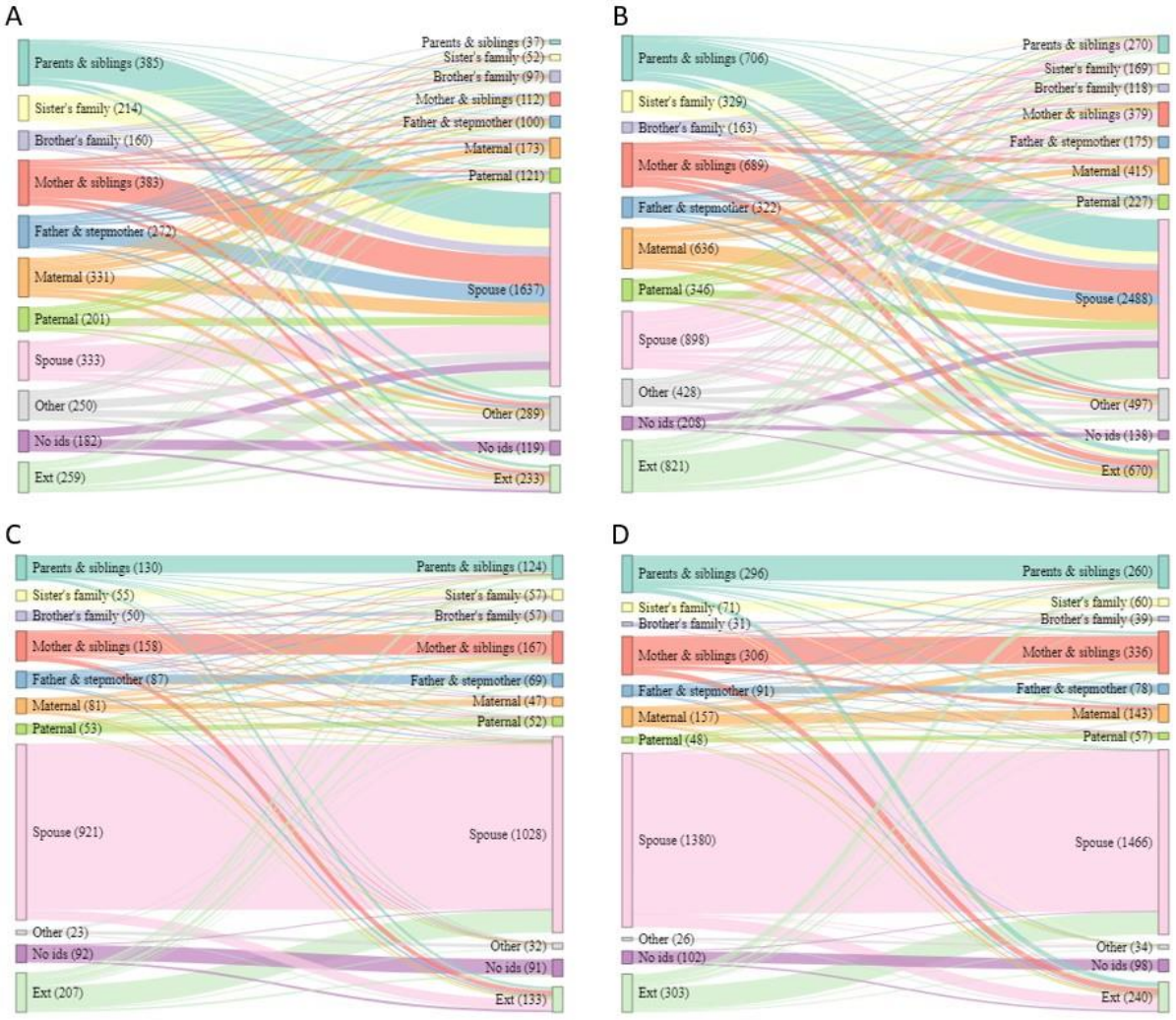
Sankey diagrams showing the household composition of sending and receiving households shown in Figures 8.3 and 8.4 show the difference in types of short move for children and adolescents (respectively). For children moving independently, there was not one more common move type, and children mostly changed household type when they moved (Figure 8.3A & B). By sex, there was some evidence that female children were more likely to move to or from '*parents and siblings'*, to '*sister's family*' or '*maternal'* or from outside of the area, while male children were more likely to move to or from '*brother's family'*, '*father & stepmother'* and '*paternal'* (Table 8.1). For accompanied children, moving from a '*parents and siblings'* household to another of the same type was most common, moving to and from '*mother and siblings'* and '*maternal'* households was also common (Figure 8.3C & D). By sex, there was some evidence that male children were more likely to move to or from '*sister's family'*, '*brother's family'* and '*father and stepmother'*, while female children were more likely to move to or from '*maternal'* (Table 8.1). For adolescents, the picture is different, for both male and female adolescents moving independently there was a wide variety of sending households, however the destination household were overwhelmingly '*spouse'* households. A visible difference between males and females in this group is that females moving from '*spouse'* households go to a wider variety of receiving households ('*parents and siblings'*, '*mother and siblings'* etc.) (Figure 8.4A & B). By sex, female adolescents were more likely to move to '*parents and siblings'*, '*sister's family'*, '*mother and siblings*', '*maternal'* and from or to outside of the area. Female adolescents were more likely to move from '*spouse'* households, but male adolescents were more likely to move to '*spouse'* households (Table 8.1). The accompanied moves for adolescents show a different picture again, with the majority moving from a '*spouse'* household to another of the same type (Figure 8.4C & D). By sex, female adolescents were more likely to move from and to '*parents and siblings'*, '*mother and siblings*' and '*maternal'* households, while male adolescents were more likely to move from and to '*brother's family*' (Table 8.1).

Figure 8.3. Sankey diagrams showing flow between sending and receiving households for all short moves by children



NB: Short moves are under 4 kilometres; independent moves are without a parent of any age or an adult aged 18 or over; part A shows independent moves by male children (aged <16), B independent moves by female children (aged <12); C accompanied moves by male children; D accompanied moves by female children

Figure 8.4. Sankey diagrams showing flow between sending and receiving households for all short moves by adolescents



NB: Short moves are under 4 kilometres; independent moves are without a parent of any age or an adult aged 18 or over; part A shows independent moves by male adolescents (aged 16-28), B independent moves by female adolescents (aged 12-24); C accompanied moves by male adolescents; D accompanied moves by female adolescents

Table 8.1. Sending or receiving household (HH) type by move type, age and sex, short moves only

	Sending HH: Female			Sending HH: Male			p	Receiving HH: Female			Receiving HH: Male			p
	n	%	95% CI	n	%	95% CI		n	%	95% CI	n	%	95% CI	
Child independent														
Par& sib	333	14.4%	12.7-16.1%	314	11.9%	10.4-13.5%	0.011	227	9.8%	8.5-11.1%	193	7.3%	6.1-8.5%	0.003
Sis family	41	1.8%	1.2-2.3%	39	1.5%	1.0-2.0%	0.432	56	2.4%	1.8-3.1%	39	1.5%	1.0-2.0%	0.026
Bro family	14	0.6%	0.3-0.9%	31	1.2%	0.7-1.6%	0.014	15	0.6%	0.3-1.0%	47	1.8%	1.2-2.3%	<0.001
Moth& sib	360	15.6%	13.8-17.3%	414	15.7%	14.0-17.4%	0.895	302	13.1%	11.5-14.6%	347	13.2%	11.7-14.6%	0.912
Fath&stmo	220	9.5%	8.1-10.9%	364	13.8%	12.2-15.4%	<0.001	315	13.6%	12.0-15.2%	473	17.9%	16.2-19.7%	<0.001
Maternal	534	23.1%	21.1-25.1%	642	24.3%	22.4-26.3%	0.31	614	26.5%	24.5-28.6%	606	23.0%	21.2-24.8%	0.004
Paternal	183	7.9%	6.7-9.2%	248	9.4%	8.1-10.7%	0.067	370	16.0%	14.3-17.7%	497	18.8%	17.1-20.6%	0.011
Other	110	4.8%	3.8-5.7%	111	4.2%	3.4-5.0%	0.367	125	5.4%	4.4-6.4%	145	5.5%	4.5-6.5%	0.889
No ids	33	1.4%	0.9-2.0%	43	1.6%	1.1-2.1%	0.586	47	2.0%	1.4-2.7%	56	2.1%	1.5-2.7%	0.832
Ext	486	21.0%	19.3-22.7%	431	16.3%	14.8-17.9%	<0.001	243	10.5%	9.2-11.8%	234	8.9%	7.7-10.0%	0.045
Child accompanied														
Par& sib	2907	53.4%	51.5-55.3%	3157	51.2%	49.4-53.1%	0.044	2472	45.4%	43.5-47.3%	2743	44.5%	42.6-46.4%	0.407
Sis family	57	1.0%	0.7-1.4%	89	1.4%	1.0-1.9%	0.056	42	0.8%	0.5-1.1%	82	1.3%	0.9-1.8%	0.005
Bro family	11	0.2%	0.1-0.3%	33	0.5%	0.3-0.8%	0.009	10	0.2%	0.1-0.3%	36	0.6%	0.3-0.8%	0.001
Moth& sib	866	15.9%	14.5-17.3%	1054	17.1%	15.8-18.5%	0.133	1246	22.9%	21.3-24.4%	1407	22.8%	21.3-24.4%	0.962
Fath&stmo	112	2.1%	1.6-2.5%	194	3.1%	2.6-3.7%	0.001	112	2.1%	1.6-2.5%	210	3.4%	2.8-4.0%	<0.001
Maternal	813	14.9%	13.6-16.3%	798	13.0%	11.8-14.1%	0.005	859	15.8%	14.5-17.0%	870	14.1%	13.0-15.3%	0.02
Paternal	42	0.8%	0.5-1.0%	67	1.1%	0.7-1.5%	0.167	50	0.9%	0.6-1.2%	70	1.1%	0.8-1.5%	0.334
Other	33	0.6%	0.3-0.9%	29	0.5%	0.3-0.7%	0.422	29	0.5%	0.3-0.8%	29	0.5%	0.3-0.6%	0.685
No ids	51	0.9%	0.7-1.2%	87	1.4%	1.0-1.8%	0.044	53	1.0%	0.7-1.3%	89	1.4%	1.0-1.8%	0.057
Ext	554	10.2%	9.2-11.2%	653	10.6%	9.6-11.6%	0.443	573	10.5%	9.5-11.6%	625	10.1%	9.1-11.2%	0.504

	Sending HH: Female			Sending HH: Male			p	Receiving HH: Female			Receiving HH: Male			p
	n	%	95% CI	n	%	95% CI		n	%	95% CI	n	%	95% CI	
Adolescent independent														
Par& sib	706	12.7%	11.7-13.8%	385	13.0%	11.6-14.3%	0.767	270	4.9%	4.3-5.4%	37	1.2%	0.8-1.7%	<0.001
Sis family	329	5.9%	5.2-6.7%	214	7.2%	6.1-8.3%	0.036	169	3.0%	2.6-3.5%	52	1.8%	1.3-2.2%	<0.001
Bro family	163	2.9%	2.4-3.4%	160	5.4%	4.5-6.3%	<0.001	118	2.1%	1.7-2.5%	97	3.3%	2.6-3.9%	0.003
Moth& sib	689	12.4%	11.4-13.5%	383	12.9%	11.5-14.3%	0.561	379	6.8%	6.1-7.5%	112	3.8%	3.1-4.5%	<0.001
Fath&stmo	322	5.8%	5.1-6.5%	272	9.2%	8.0-10.3%	<0.001	175	3.2%	2.7-3.6%	100	3.4%	2.7-4.0%	0.599
Maternal	636	11.5%	10.5-12.4%	331	11.1%	9.9-12.4%	0.668	415	7.5%	6.8-8.2%	173	5.8%	4.9-6.7%	0.003
Paternal	346	6.2%	5.5-7.0%	201	6.8%	5.7-7.8%	0.381	227	4.1%	3.6-4.6%	121	4.1%	3.3-4.8%	0.968
Spouse	898	16.2%	15.2-17.2%	333	11.2%	10.0-12.4%	<0.001	2488	44.9%	43.5-46.2%	1637	55.1%	53.2-57.0%	<0.001
Other	428	7.7%	7.0-8.5%	250	8.4%	7.4-9.5%	0.28	497	9.0%	8.2-9.7%	289	9.7%	8.6-10.8%	0.267
No ids	208	3.8%	3.2-4.3%	182	6.1%	5.2-7.0%	<0.001	138	2.5%	2.1-2.9%	119	4.0%	3.2-4.8%	0.001
Ext	821	14.8%	13.9-15.7%	259	8.7%	7.7-9.7%	<0.001	670	12.1%	11.2-13.0%	233	7.8%	6.9-8.8%	<0.001
Adolescent accompanied														
Par& sib	296	10.5%	9.1-12.0%	130	7.0%	5.5-8.5%	<0.001	260	9.2%	7.9-10.6%	124	6.7%	5.2-8.1%	0.002
Sis family	71	2.5%	1.8-3.2%	55	3.0%	2.0-3.9%	0.418	60	2.1%	1.4-2.8%	57	3.1%	2.1-4.1%	0.107
Bro family	31	1.1%	0.6-1.6%	50	2.7%	1.8-3.6%	0.002	39	1.4%	0.9-1.9%	57	3.1%	2.2-4.0%	0.001
Moth& sib	306	10.9%	9.5-12.3%	158	8.5%	6.9-10.1%	0.011	336	12.0%	10.5-13.4%	167	9.0%	7.5-10.5%	0.001
Fath&stmo	91	3.2%	2.4-4.1%	87	4.7%	3.4-6.0%	0.04	78	2.8%	2.0-3.5%	69	3.7%	2.6-4.8%	0.115
Maternal	157	5.6%	4.6-6.6%	81	4.4%	3.3-5.5%	0.082	143	5.1%	4.1-6.0%	47	2.5%	1.7-3.3%	<0.001
Paternal	48	1.7%	1.2-2.3%	53	2.9%	2.0-3.8%	0.027	57	2.0%	1.4-2.6%	52	2.8%	2.0-3.6%	0.129
Spouse	1380	49.1%	46.8-51.4%	921	49.6%	46.9-52.3%	0.703	1466	52.2%	49.9-54.4%	1028	55.4%	52.6-58.1%	0.02
Other	26	0.9%	0.5-1.4%	23	1.2%	0.6-1.9%	0.444	34	1.2%	0.7-1.7%	32	1.7%	0.9-2.5%	0.267
No ids	102	3.6%	2.9-4.4%	92	5.0%	3.8-6.2%	0.054	98	3.5%	2.7-4.3%	91	4.9%	3.7-6.1%	0.044
Ext	303	10.8%	9.5-12.0%	207	11.1%	9.6-12.7%	0.651	240	8.5%	7.4-9.7%	133	7.2%	5.8-8.5%	0.049

8.4.4. Moving with parents

The proportion of accompanied movers according to whether they moved with their parents is shown in Table 8.2. Males tended to be more likely than females to move with neither parent; females were more likely to move with just mother, and male children with just their father (though this was not common); female adolescents were more likely to move with both parents for short moves.

Table 8.2. Whether moved with parents by move type (accompanied only), length and sex

	Female				Male				p
	n	%	95% CI		n	%	95% CI		
SHORT MOVES									
Child accompanied									
Neither	323	5.9%	5.2%	6.7%	454	7.4%	6.5%	8.2%	0.005
Mother only	2652	48.7%	46.7%	50.6%	2831	46.0%	44.0%	47.9%	0.011
Father only	121	2.2%	1.7%	2.7%	223	3.6%	3.0%	4.3%	<0.001
Both	2350	43.2%	41.2%	45.1%	2653	43.1%	41.1%	45.0%	0.934
Adolescent accompanied									
Neither	1947	69.3%	67.1%	71.4%	1425	76.7%	74.3%	79.2%	<0.001
Mother only	469	16.7%	15.0%	18.4%	217	11.7%	9.9%	13.5%	<0.001
Father only	80	2.8%	2.1%	3.6%	62	3.3%	2.3%	4.4%	0.407
Both	315	11.2%	9.6%	12.8%	153	8.2%	6.6%	9.9%	0.002
LONG MOVES									
Child accompanied									
Neither	816	8.7%	8.0%	9.4%	997	9.4%	8.7%	10.1%	0.132
Mother only	4145	44.3%	42.8%	45.8%	4368	41.2%	39.8%	42.6%	<0.001
Father only	262	2.8%	2.4%	3.2%	401	3.8%	3.3%	4.3%	0.001
Both	4133	44.2%	42.7%	45.7%	4836	45.6%	44.1%	47.1%	0.088
Adolescent accompanied									
Neither	3914	73.8%	72.1%	75.4%	2636	78.1%	76.3%	80.0%	<0.001
Mother only	534	10.1%	9.0%	11.1%	222	6.6%	5.5%	7.7%	<0.001
Father only	141	2.7%	2.1%	3.2%	111	3.3%	2.5%	4.1%	0.149
Both	716	13.5%	12.1%	14.8%	404	12.0%	10.4%	13.5%	0.056

NB: Short moves are under 4 kilometres; independent moves are without a parent of any age or an adult aged 18 or over; female children are aged <12, male children are aged <16, female adolescents are aged 12-24 and male adolescents are aged 16-28; confidence intervals and p-values calculated allowing for clustering by unique household ID and unique individual ID

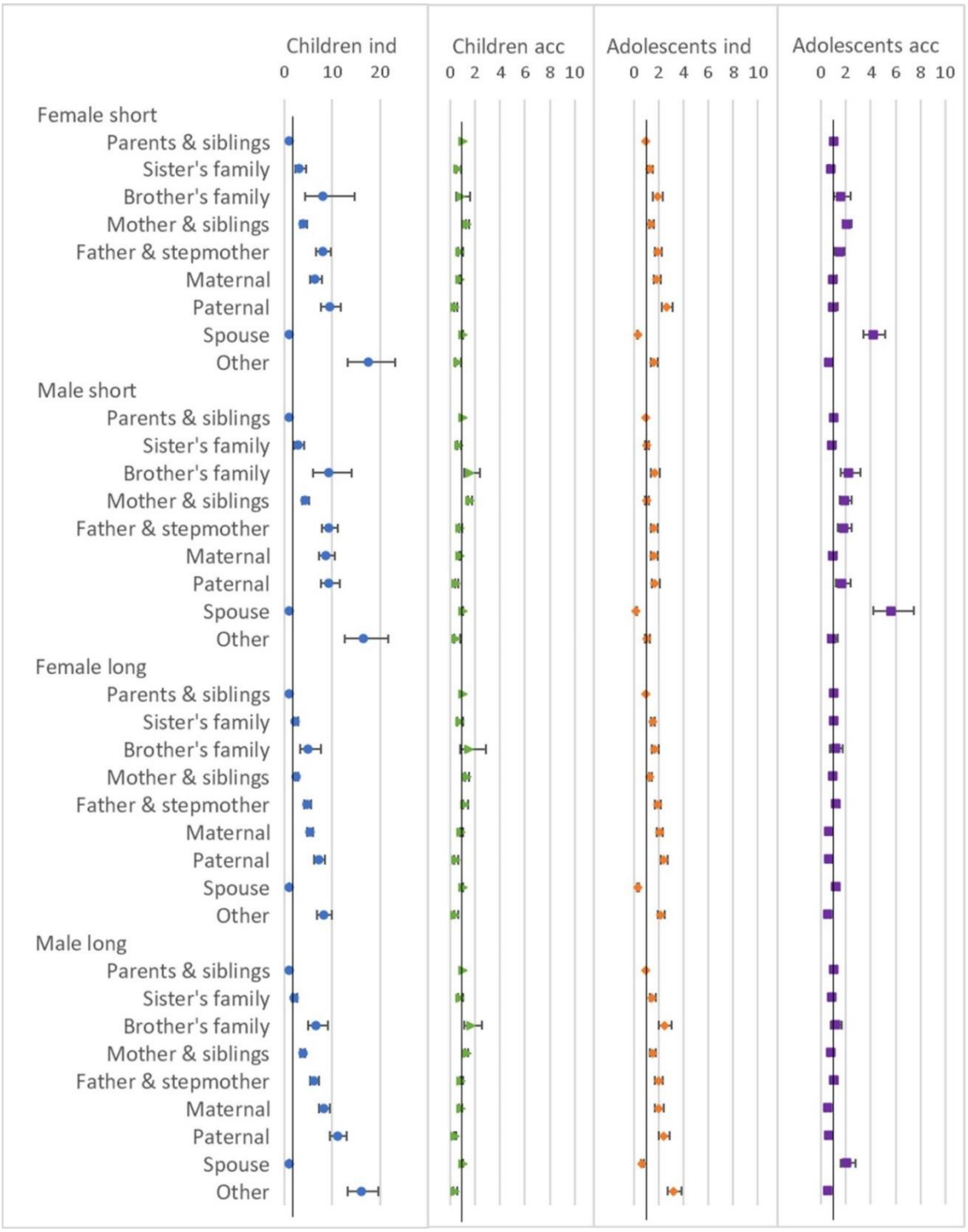
8.4.5. Regression analysis of association of family and household composition/structure with mobility

A total of 7783 (1153 female children, 1034 male children, 3679 female adolescents and 1917 male adolescents) were dropped from the regression analysis as they had no parent identifiers, so family variables could not be constructed; 36,128 moves were dropped as well as they were moves from outside of the HDSS so did not have information on the sending household/area or because the mover did not have parent IDs (Figure 8.1). The full results from the models, including confounding variables, are displayed in the tables 8.3-8.6, while figures displaying the key results from the three family and household composition/structure variables plus selected other factors (highlighted as they showed interesting associations) are discussed below.

Household composition

Results from the regression models for the sending household composition variable are shown in Figure 8.5. In all cases the baseline category was '*parents and siblings'*. While there are associations between this variable and mobility for all age groups and types of move, the associations are strongest for children moving independently. For this group, the highest odds ratios tended to be for the '*other'* category (meaning children living in this category were more likely to move out than those living with *'parents and siblings'*), followed by '*paternal'*, '*brother's family'*, '*maternal'* and '*father and stepmother*'. For child accompanied moves, the odds ratios were closer to 1 than for the independent moves, suggesting the composition of the sending household was less strongly associated with mobility for accompanied than independent moves, and the associations were more varied: '*sister's family'*, '*paternal'* and '*other'* were associated with slightly lower odds of moving and '*mother and siblings'* with slightly higher odds. The odds ratios for adolescents moving independently followed a similar pattern to those of the children, albeit closer to 1. The '*spouse'* household type, which was not present for children, is the only category associated with lower odds of move, suggesting adolescents were less likely to leave *'spouse'* households independently than *'parents and siblings'* household. The odds ratios for adolescent accompanied moves tended to be closer to 1, and somewhat similar to the patterns seen for children; for this move type the *'spouse'* category was mostly strongly associated with higher odds of moving, for all except female long moves.

Figure 8.5. Odds ratios relating to sending household composition from multinomial multi-level regression models

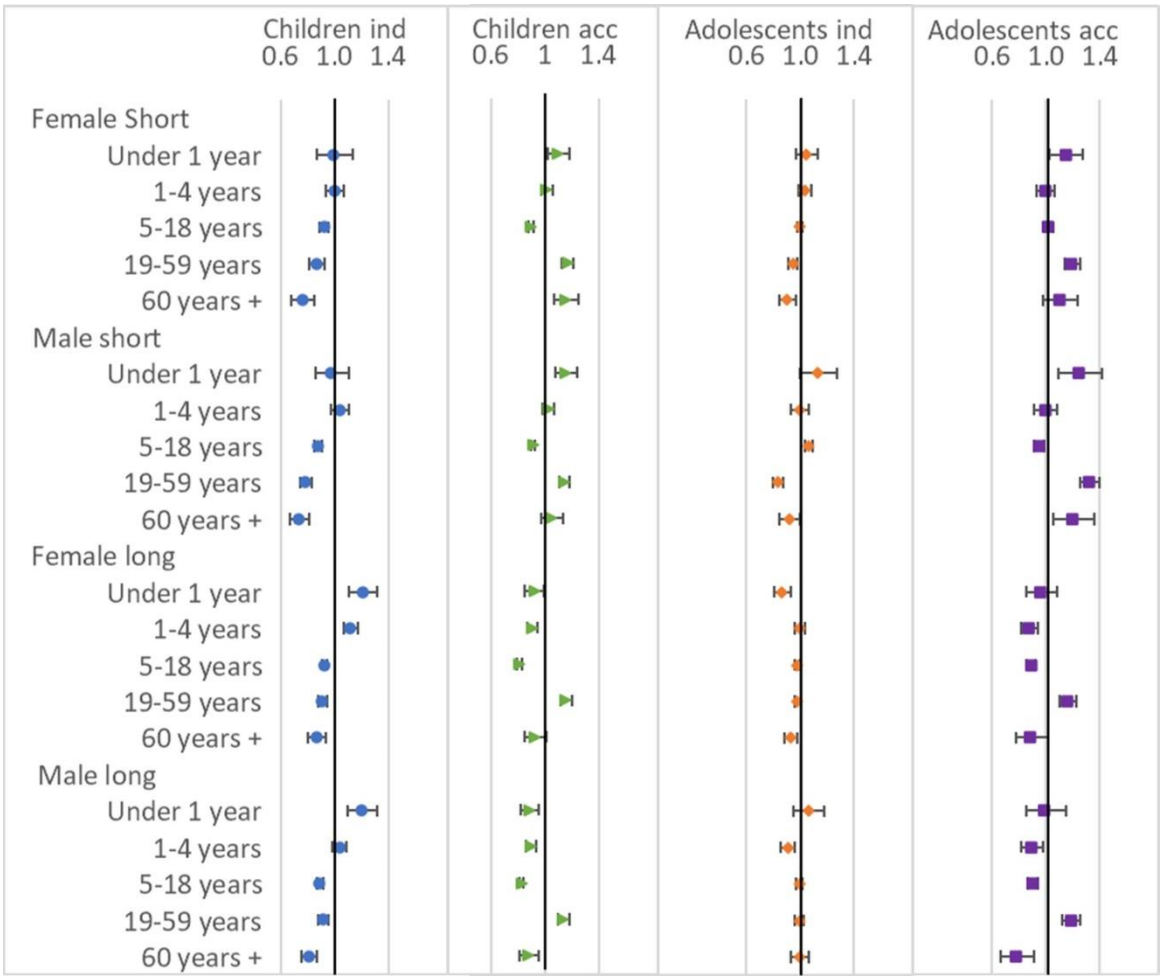


NB: The models run separately for female children (aged <12), male children (aged <16), female adolescents (aged 12-24) and male adolescents (aged 16-28); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id; household composition is a single categorical variable with ‘parents & siblings’ as baseline; models control for presence of own child (adolescents only), household age composition, presence of family within 250 metres, orphanhood, age, year, distance to tarmac road, population density, parental education and household head employment score; note the differing axes for ‘children ind’ compared to the others

Household age composition

Results from the regression models for the household age composition variables are shown in Figure 8.6. For children moving independently, presence of increasing number of children aged under 5 in the household either has no association or was associated with higher chance of moving; while increasing number of people aged five and over was associated with lower chance of move, with the odds ratios for the number of people aged 60 years+ the furthest from 1 (suggesting the lowest chance of moving). The effects were different for child accompanied moves, with increasing number of children aged 18 and under having no association or associated with a lower chance of move; increasing number of 19–59-year-olds was associated with higher odds of move; and increasing number of adults aged 60 or over was associated with higher odds for short move but lower odds for long moves. For adolescents moving independently there was not such a clear pattern, but increasing numbers of other household members tended to be associated with lower odds of move: for long moves presence of children aged under one year (but not older children) was associated with lower chance of move for females; however for males this association was seen with one-four year-olds (but not younger children). For accompanied adolescents, increasing numbers of 19–59-year-olds was consistently associated with higher chance of move, while, for long moves, all other age groups (children and those aged 60+) were associated with lower chance of move.

Figure 8.6. Odds ratios relating to sending household age composition from multinomial multi-level regression models



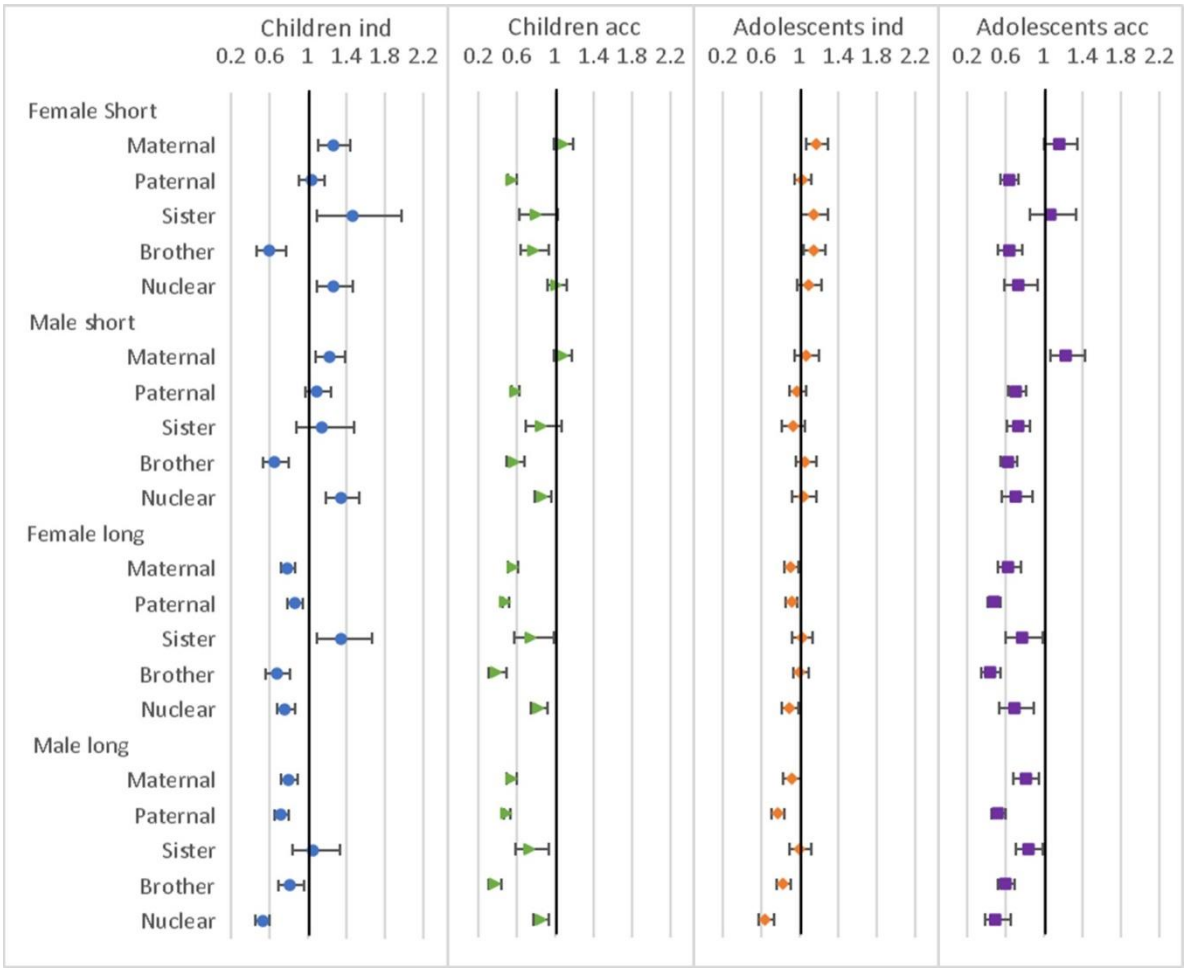
NB: The models run separately for female children (aged <12), male children (aged <16), female adolescents (aged 12-24) and male adolescents (aged 16-28); outcome is move type with 'no move' as baseline; models allow for clustering by unique household id and unique participant id; household age composition are all individual variables of total number in each age group [excluding index] in the household; models control for household composition, presence of own child (adolescents only), presence of family within 250 metres, orphanhood, age, year, distance to tarmac road, population density, parental education and household head employment score;

Individual relative variables

Results from the regression models for the presence of family within 250 metres (but not in the household) variables are shown in Figure 8.7. For children moving independently, presence of maternal and nuclear family (and sister's family for female children only) were associated with higher chance of move, while brother's family was associated with a lower chance. For long moves the pattern was different with all family types associated with lower chance of move, except sister, which was associated with higher chance for females. For accompanied moves, there was either an association with lower chance of move for all family types, or, in few cases, no association. For the adolescents moving independently, the

odds ratios tended to be closer to 1, particularly for short moves, but the pattern for male long moves was similar to that of the children. For accompanied moves most family types were associated with lower chance of move, except maternal for short moves.

Figure 8.7. Odds ratios relating to family living nearby from multinomial multi-level regression models



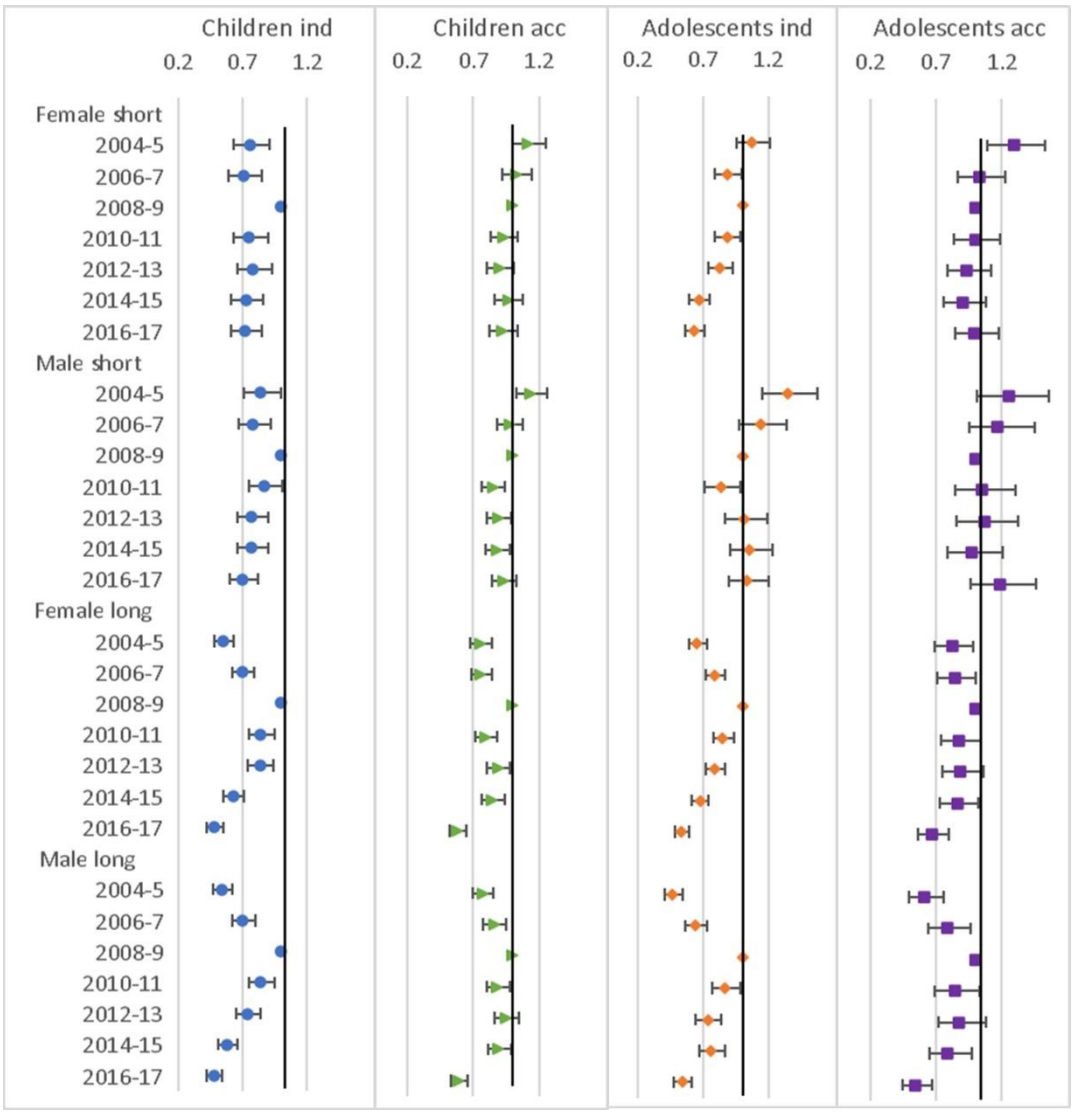
NB: The models run separately for female children (aged <12), male children (aged <16), female adolescents (aged 12-24) and male adolescents (aged 16-28); outcome is move type with 'no move' as baseline; models allow for clustering by unique household id and unique participant id; family nearby are all individual binary variables; models control for household composition, presence of own child (adolescents only), household age composition, orphanhood, age, year, distance to tarmac road, population density, parental education and household head employment score

Other factors

The effect of calendar year is clearest for the long moves for both children and adolescents and short and long moves, where there was less chance of moving both before and after the baseline period of 2008-9, and the odds of moving getting smaller with later calendar time. For short moves, there was a similar pattern (though closer to 1) for children moving independently but for the other groups it was less clear (Figure 8.8). The employment level

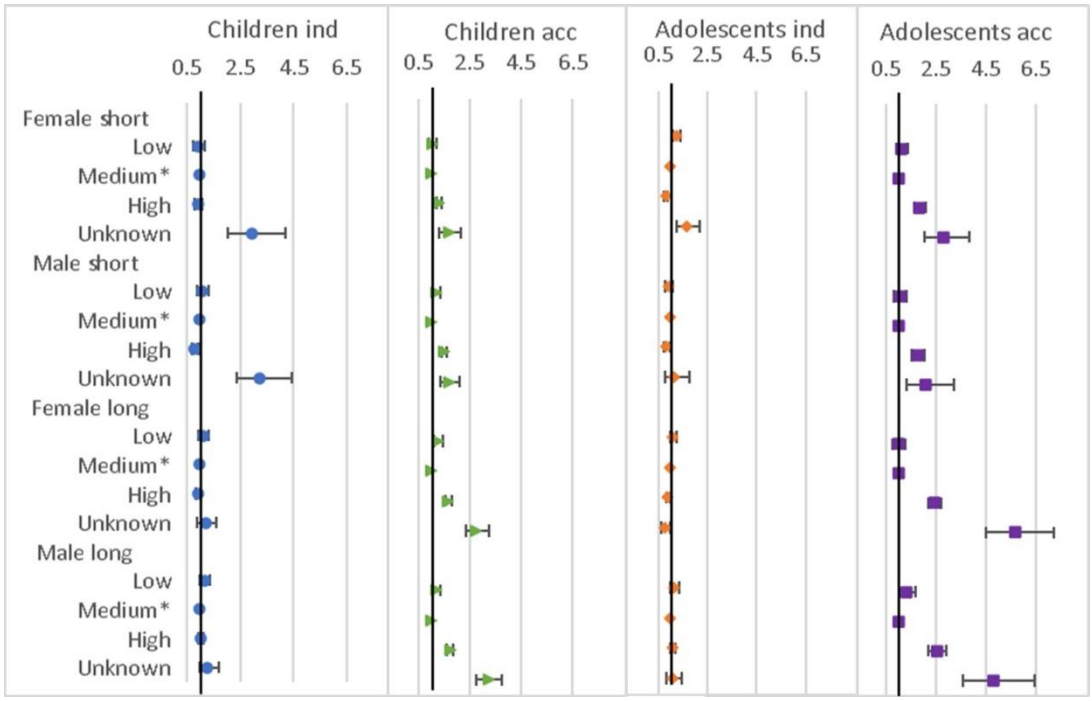
of the household head of the sending household had little effect on independent moves, though for short independent child moves those with unknown status were more likely to move. For accompanied moves those in the high and unknown category had higher odds of moving compared to the medium category, this was particularly the case for long moves and adolescents (Figure 8.9).

Figure 8.8. Odds ratios relating to calendar year from multinomial multi-level regression models



NB: The models run separately for female children (aged <12), male children (aged <16), female adolescents (aged 12-24) and male adolescents (aged 16-28); outcome is move type with 'no move' as baseline; models allow for clustering by unique household id and unique participant id; calendar year is in 2-year categories with 2008-9 as baseline; models control for household composition, presence of own child (adolescents only), household age composition, presence of family within 250 metres, orphanhood, age, distance to tarmac road, population density, parental education and household head employment score

Figure 8.9. Odds ratios relating to household head employment score from multinomial multi-level regression models



NB: The models run separately for female children (aged <12), male children (aged <16), female adolescents (aged 12-24) and male adolescents (aged 16-28); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id; household head employment score is in 3 categories with ‘medium’ as baseline; models control for household composition, presence of own child (adolescents only), household age composition, presence of family within 250 metres, orphanhood, age, year, distance to tarmac road, population density and parental education

Table 8.3a. Results for short independent outcome for children, from multi-level multinomial regression models

	Children (short independent moves)											
	Female (1795 moves)						Male (2163 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Household composition												
Par & siblings	Reference						Reference					
Sister's family	3.1	2.2	4.5	0.18	6.3	<0.001	2.9	2.0	4.1	0.18	5.9	<0.001
Brother's family	8.0	4.4	14.6	0.31	6.8	<0.001	9.2	6.0	14.1	0.22	10.2	<0.001
Moth & siblings	4.0	3.3	4.7	0.09	15.2	<0.001	4.3	3.6	5.1	0.09	16.7	<0.001
Fath & stepmoth	8.0	6.6	9.8	0.10	20.6	<0.001	9.4	7.9	11.1	0.09	25.3	<0.001
Maternal	6.4	5.3	7.8	0.10	19.3	<0.001	8.7	7.3	10.5	0.09	23.3	<0.001
Paternal	9.5	7.6	11.9	0.11	19.9	<0.001	9.4	7.6	11.6	0.11	21.1	<0.001
Spouse	1.0	1.0	1.0	0.00		<0.001	1.0	1.0	1.0	0.00		<0.001
Other	17.5	13.3	23.1	0.14	20.4	<0.001	16.5	12.6	21.6	0.14	20.4	<0.001
Own child (bin)	1.0	1.0	1.0			<0.001	1.0	1.0	1.0			<0.001
Presence in household (all individual continuous variables)												
Under 1 year	1.0	0.9	1.1	0.07	-0.1	0.934	1.0	0.9	1.1	0.07	-0.4	0.704
1-4 years	1.0	0.9	1.1	0.04	0.0	0.961	1.0	1.0	1.1	0.03	1.2	0.226
5-18 years	0.9	0.9	1.0	0.02	-4.8	<0.001	0.9	0.8	0.9	0.02	-8.4	<0.001
19-59 years	0.9	0.8	0.9	0.03	-4.6	<0.001	0.8	0.7	0.8	0.03	-8.5	<0.001
60 years & over	0.8	0.7	0.9	0.06	-4.8	<0.001	0.7	0.7	0.8	0.05	-6.1	<0.001
Presence of family type within 250m (all individual binary variables)												
Maternal	1.3	1.1	1.4	0.07	3.4	0.001	1.2	1.1	1.4	0.06	3.1	0.002
Paternal	1.0	0.9	1.2	0.07	0.5	0.620	1.1	1.0	1.2	0.06	1.5	0.126
Sister	1.5	1.1	2.0	0.15	2.6	0.010	1.1	0.9	1.5	0.13	1.0	0.309
Brother	0.6	0.5	0.8	0.13	-3.9	<0.001	0.7	0.5	0.8	0.10	-4.2	<0.001
Nuclear	1.3	1.1	1.5	0.08	3.1	0.002	1.3	1.2	1.5	0.07	4.6	<0.001
Mat orphan (bin)	1.2	0.9	1.6	0.15	1.1	0.282	1.4	1.1	1.8	0.12	2.9	0.004
Pat orphan (bin)	1.0	0.8	1.2	0.10	-0.2	0.870	1.1	1.0	1.3	0.08	1.6	0.109
Age (cont.)	1.1	1.1	1.1	0.01	7.9	<0.001	1.0	1.0	1.0	0.01	4.3	<0.001
Calendar year (categorical)												
2004-5	0.8	0.6	0.9	0.10	-2.9	0.004	0.8	0.7	1.0	0.09	-2.0	0.042
2006-7	0.7	0.6	0.8	0.09	-3.8	<0.001	0.8	0.7	0.9	0.08	-3.0	0.003
2008-9	Reference						Reference					
2010-11	0.8	0.6	0.9	0.09	-3.2	0.001	0.9	0.7	1.0	0.08	-1.8	0.071
2012-13	0.8	0.7	0.9	0.09	-2.8	0.005	0.8	0.7	0.9	0.08	-3.3	0.001
2014-15	0.7	0.6	0.9	0.09	-3.7	<0.001	0.8	0.7	0.9	0.08	-3.4	0.001
2016-17	0.7	0.6	0.9	0.09	-3.8	<0.001	0.7	0.6	0.8	0.08	-4.5	<0.001

	Children (short independent moves)											
	Female (1795 moves)						Male (2163 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Distance to tarmac road in km (cont integer)	1.0	1.0	1.0	0.01	-2.7	0.008	1.0	1.0	1.0	0.01	-2.5	0.011
Population density within 250m (categorical)												
<50p	1.2	1.0	1.3	0.08	2.0	0.045	1.1	1.0	1.3	0.07	1.8	0.064
50-149p	Reference						Reference					
150-299p	1.2	1.0	1.3	0.07	2.1	0.038	1.1	0.9	1.2	0.06	1.1	0.274
>=300p	1.0	0.9	1.2	0.08	0.1	0.948	0.9	0.8	1.1	0.08	-0.8	0.412
Father secondary education (categorical)												
None	Reference						Reference					
Secondary	0.9	0.8	1.1	0.06	-0.8	0.398	1.1	1.0	1.2	0.06	1.5	0.143
Unknown	0.9	0.6	1.4	0.21	-0.3	0.762	0.9	0.6	1.3	0.21	-0.6	0.559
Mother secondary education (categorical)												
None	Reference						Reference					
Secondary	1.0	0.9	1.2	0.06	0.4	0.662	1.0	0.9	1.1	0.06	-0.1	0.940
Unknown	0.4	0.2	0.8	0.29	-2.9	0.003	0.5	0.3	0.9	0.26	-2.4	0.018
Household head employment score (categorical)												
Low	0.9	0.8	1.2	0.12	-0.5	0.646	1.1	0.9	1.3	0.10	1.0	0.330
Medium	Reference						Reference					
High	0.9	0.8	1.1	0.08	-1.0	0.334	0.8	0.7	1.0	0.07	-2.6	0.009
Unknown	2.9	2.1	4.2	0.18	6.0	<0.001	3.3	2.4	4.4	0.16	7.5	<0.001

NB: Models run separately for female children (aged <12 [19,650 individuals]) and male children (aged <16 [21,267 individuals]); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id; short moves are under 4 kilometres; independent moves are without a parent of any age or an adult aged 18 or over

Table 8.3b. Results for long independent move outcome for children, from multi-level multinomial regression models

	Children (long independent moves)											
	Female (3425 moves)						Male (3321 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Household composition												
Par & siblings	Reference						Reference					
Sister's family	2.4	1.9	2.9	0.11	8	<0.001	2.1	1.7	2.7	0.13	5.9	<0.001
Broth family	5	3.2	7.6	0.22	7.3	<0.001	6.7	4.9	9.2	0.16	11.8	<0.001
Moth & sibl	2.5	2.2	2.8	0.07	14.1	<0.001	3.9	3.4	4.4	0.07	19.6	<0.001
Fath & stepmo	4.8	4.1	5.5	0.07	21.3	<0.001	6.3	5.4	7.3	0.07	25	<0.001
Maternal	5.3	4.7	6	0.06	25.9	<0.001	8.3	7.3	9.6	0.07	30.2	<0.001
Paternal	7.2	6.2	8.4	0.08	26	<0.001	11.2	9.6	13.1	0.08	30.3	<0.001
Other	8.2	6.8	10	0.1	21.3	<0.001	16.2	13.3	19.6	0.1	28.3	<0.001
Presence in household (all individual continuous variables)												
Under 1 year	1.2	1.1	1.3	0.04	4.3	<0.001	1.2	1.1	1.3	0.05	3.8	<0.001
1-4 years	1.1	1.1	1.2	0.02	4.6	<0.001	1.0	1.0	1.1	0.03	1.4	0.164
5-18 years	0.9	0.9	0.9	0.01	-6.7	<0.001	0.9	0.9	0.9	0.01	-9.5	<0.001
19-59 years	0.9	0.9	0.9	0.02	-4.5	<0.001	0.9	0.9	1.0	0.02	-4.4	<0.001
60 years +	0.9	0.8	0.9	0.04	-3.8	<0.001	0.8	0.8	0.9	0.04	-5.8	<0.001
Presence of family type within 250m (all individual binary variables)												
Maternal	0.8	0.7	0.9	0.05	-4.7	<0.001	0.8	0.7	0.9	0.05	-4.3	<0.001
Paternal	0.9	0.8	0.9	0.05	-3.2	0.002	0.7	0.7	0.8	0.05	-6.6	<0.001
Sister	1.4	1.1	1.7	0.11	2.8	0.006	1.1	0.8	1.3	0.12	0.5	0.648
Brother	0.7	0.6	0.8	0.09	-4.2	<0.001	0.8	0.7	1.0	0.08	-2.5	0.013
Nuclear	0.8	0.7	0.9	0.06	-4.3	<0.001	0.5	0.5	0.6	0.07	-9.1	<0.001
Mat orph (bin)	1.2	1.0	1.5	0.1	1.8	0.075	1.1	0.9	1.3	0.09	1.2	0.229
Pat orph (bin)	0.9	0.8	1.0	0.07	-1.5	0.133	0.9	0.8	1.0	0.06	-2.1	0.039
Age (cont)	1.2	1.2	1.2	0.01	25.9	<0.001	1.1	1.1	1.1	0.01	16.9	<0.001
Calendar year (categorical)												
2004-5	0.6	0.5	0.6	0.07	-8.6	<0.001	0.5	0.5	0.6	0.07	-8.7	<0.001
2006-7	0.7	0.6	0.8	0.06	-5.7	<0.001	0.7	0.6	0.8	0.06	-5.6	<0.001
2008-9	Reference						Reference					
2010-11	0.8	0.7	1.0	0.06	-2.8	0.005	0.8	0.7	0.9	0.06	-2.8	0.005
2012-13	0.8	0.7	0.9	0.06	-2.9	0.003	0.7	0.7	0.8	0.06	-4.8	<0.001
2014-15	0.6	0.6	0.7	0.06	-7.2	<0.001	0.6	0.5	0.7	0.07	-8.4	<0.001
2016-17	0.5	0.4	0.6	0.07	-10.8	<0.001	0.5	0.4	0.5	0.07	-11	<0.001
Distance to tarmac road in km (cont int)	1.0	1.0	1.0	0.01	0.7	0.472	1.0	1.0	1.0	0.01	0.2	0.809

	Children (long independent moves)											
	Female (3425 moves)						Male (3321 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Population density within 250m (categorical)												
<50p	1.1	1.0	1.2	0.05	2.2	0.025	1.1	1.0	1.3	0.05	2.6	0.009
50-149p	Reference						Reference					
150-299p	1.1	1.0	1.2	0.05	1.2	0.238	1.2	1.1	1.3	0.05	3	0.003
>=300p	1.1	1.0	1.2	0.06	1.9	0.059	1.1	1.0	1.2	0.06	1.5	0.134
Father secondary education (categorical)												
None	Reference						Reference					
Secondary	1.0	1.0	1.1	0.04	1.0	0.339	0.9	0.8	1.0	0.05	-1.9	0.052
Unknown	3.5	2.9	4.3	0.1	12.5	<0.001	3.3	2.7	4.0	0.1	11.5	<0.001
Mother secondary education (categorical)												
None	Reference						Reference					
Secondary	1.0	0.9	1.1	0.05	-0.9	0.387	1.0	0.9	1.1	0.05	0.7	0.5
Unknown	1.0	0.8	1.2	0.12	-0.3	0.745	1.1	0.9	1.4	0.12	0.8	0.399
Household head employment score (categorical)												
Low	1.1	1.0	1.3	0.08	1.6	0.108	1.2	1.0	1.4	0.08	2.2	0.027
Medium	Reference						Reference					
High	1.0	0.9	1.1	0.05	-0.9	0.346	1.0	0.9	1.1	0.05	0.5	0.591
Unknown	1.2	0.9	1.6	0.15	1.3	0.202	1.3	1.0	1.7	0.14	1.9	0.055

NB: Models run separately for female children (aged <12 [19,650 individuals]) and male children (aged <16 [21,267 individuals]); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id

Table 8.4a. Results for short accompanied and long accompanied move outcome for children, from multi-level multinomial regression models

	Children (short accompanied moves)											
	Female (4841 moves)						Male (5421 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Household composition												
Par & siblings	Reference						Reference					
Sister's family	0.7	0.5	0.9	0.14	-3.0	0.003	0.7	0.6	0.9	0.12	-2.8	0.004
Brother's family	0.8	0.5	1.6	0.32	-0.5	0.611	1.6	1.1	2.4	0.20	2.4	0.015
Mother & siblings	1.4	1.3	1.5	0.05	6.6	<0.001	1.6	1.4	1.7	0.05	9.9	<0.001
Fath & stepmoth	0.8	0.7	1.0	0.11	-1.8	0.076	0.8	0.7	1.0	0.08	-2.3	0.022
Maternal	0.8	0.7	0.9	0.06	-4.1	<0.001	0.8	0.7	0.9	0.06	-4.5	<0.001
Paternal	0.4	0.3	0.6	0.16	-5.5	<0.001	0.5	0.4	0.6	0.13	-5.7	<0.001
Other	0.6	0.4	0.9	0.19	-2.7	0.007	0.5	0.4	0.8	0.20	-3.4	0.001
Presence in household (all individual continuous variables)												
Under 1 year	1.1	1.0	1.2	0.04	2.4	0.014	1.2	1.1	1.2	0.04	4.1	<0.001
1-4 years	1.0	1.0	1.1	0.02	0.6	0.554	1.0	1.0	1.1	0.02	1.2	0.237
5-18 years	0.9	0.9	0.9	0.01	-10.4	<0.001	0.9	0.9	0.9	0.01	-9.2	<0.001
19-59 years	1.2	1.1	1.2	0.02	8.6	<0.001	1.1	1.1	1.2	0.02	7.6	<0.001
60 years & over	1.2	1.1	1.2	0.04	3.5	<0.001	1.1	1.0	1.1	0.04	1.3	0.206
Maternal	1.1	1.0	1.2	0.05	1.8	0.064	1.1	1.0	1.2	0.04	1.9	0.063
Paternal	0.6	0.5	0.6	0.04	-15.2	<0.001	0.6	0.5	0.6	0.04	-14.9	<0.001
Sister	0.8	0.6	1.0	0.13	-1.7	0.092	0.9	0.7	1.1	0.11	-1.3	0.181
Brother	0.8	0.7	0.9	0.09	-2.7	0.007	0.6	0.5	0.7	0.08	-6.6	<0.001
Nuclear	1.0	0.9	1.1	0.05	0.4	0.717	0.9	0.8	1.0	0.05	-2.8	0.005
Mat orphan (bin)	1.2	0.9	1.7	0.17	1.3	0.183	1.3	1.0	1.7	0.14	1.7	0.088
Pat orphan (bin)	1.2	1.0	1.4	0.08	2.2	0.031	1.1	1.0	1.2	0.07	1.2	0.214
Age (continuous)	0.9	0.9	0.9	0.01	-19.5	<0.001	0.9	0.9	0.9	0.00	-22.5	<0.001
Calendar year (categorical)												
2004-5	1.1	1.0	1.2	0.06	1.9	0.052	1.1	1.0	1.3	0.05	2.5	0.012
2006-7	1.0	0.9	1.1	0.06	0.5	0.650	1.0	0.9	1.1	0.05	-0.5	0.636
2008-9	Reference						Reference					
2010-11	0.9	0.8	1.0	0.06	-1.3	0.211	0.8	0.8	0.9	0.05	-3.1	0.002
2012-13	0.9	0.8	1.0	0.06	-1.8	0.070	0.9	0.8	1.0	0.05	-2.2	0.026
2014-15	1.0	0.9	1.1	0.06	-0.6	0.527	0.9	0.8	1.0	0.05	-2.5	0.014
2016-17	0.9	0.8	1.0	0.06	-1.4	0.172	0.9	0.8	1.0	0.05	-1.4	0.167
Distance to tarmac road in km (cont. integer)	1.0	1.0	1.0	0.01	-4.6	<0.001	1.0	1.0	1.0	0.01	-3.2	0.001

	Children (short accompanied moves)											
	Female (4841 moves)						Male (5421 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Population density within 250m (categorical)												
<50p	0.9	0.8	1.0	0.05	-1.6	0.101	0.8	0.7	0.9	0.05	-4.8	<0.001
50-149p	Reference						Reference					
150-299p	1.2	1.1	1.4	0.04	5.0	<0.001	1.2	1.1	1.3	0.04	4.6	<0.001
>=300p	1.8	1.6	1.9	0.05	12.6	<0.001	1.7	1.6	1.9	0.04	12.8	<0.001
Father secondary education (categorical)												
None	Reference						Reference					
Secondary	1.1	1.0	1.2	0.04	2.2	0.027	1.0	1.0	1.1	0.04	1.2	0.248
Unknown	0.7	0.5	1.0	0.17	-2.2	0.030	0.8	0.6	1.0	0.15	-1.8	0.079
Mother secondary education (categorical)												
None	Reference						Reference					
Secondary	1.1	1.0	1.1	0.04	1.3	0.181	1.0	1.0	1.1	0.04	1.2	0.247
Unknown	1.6	1.2	2.3	0.18	2.8	0.006	1.7	1.3	2.3	0.16	3.4	0.001
Household head employment score (categorical)												
Low	1.0	0.9	1.2	0.07	0.6	0.544	1.2	1.0	1.4	0.07	2.5	0.013
Medium	Reference						Reference					
High	1.3	1.2	1.4	0.04	6.5	<0.001	1.5	1.4	1.6	0.04	10.6	<0.001
Unknown	1.7	1.3	2.1	0.12	4.2	<0.001	1.7	1.4	2.1	0.11	4.7	<0.001

NB: Models run separately for female children (aged <12 [19,650 individuals]) and male children (aged <16 [21,267 individuals]); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id

Table 8.4b. Results for long accompanied move outcome for children, from multi-level multinomial regression models

	Children (long independent moves)											
	Female (3425 moves)						Male (3321 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Household composition												
Par & siblings	Reference						Reference					
Sister's family	2.4	1.9	2.9	0.11	8.0	<0.001	2.1	1.7	2.7	0.13	5.9	<0.001
Broth family	5.0	3.2	7.6	0.22	7.3	<0.001	6.7	4.9	9.2	0.16	11.8	<0.001
Moth & sibl	2.5	2.2	2.8	0.07	14.1	<0.001	3.9	3.4	4.4	0.07	19.6	<0.001
Fath & stepmo	4.8	4.1	5.5	0.07	21.3	<0.001	6.3	5.4	7.3	0.07	25	<0.001
Maternal	5.3	4.7	6.0	0.06	25.9	<0.001	8.3	7.3	9.6	0.07	30.2	<0.001
Paternal	7.2	6.2	8.4	0.08	26	<0.001	11.2	9.6	13.1	0.08	30.3	<0.001
Other	8.2	6.8	10	0.1	21.3	<0.001	16.2	13.3	19.6	0.1	28.3	<0.001
Presence in household (all individual continuous variables)												
Under 1 year	1.2	1.1	1.3	0.04	4.3	<0.001	1.2	1.1	1.3	0.05	3.8	<0.001
1-4 years	1.1	1.1	1.2	0.02	4.6	<0.001	1.0	1.0	1.1	0.03	1.4	0.164
5-18 years	0.9	0.9	0.9	0.01	-6.7	<0.001	0.9	0.9	0.9	0.01	-9.5	<0.001
19-59 years	0.9	0.9	0.9	0.02	-4.5	<0.001	0.9	0.9	1.0	0.02	-4.4	<0.001
60 years +	0.9	0.8	0.9	0.04	-3.8	<0.001	0.8	0.8	0.9	0.04	-5.8	<0.001
Presence of family type within 250m (all individual binary variables)												
Maternal	0.8	0.7	0.9	0.05	-4.7	<0.001	0.8	0.7	0.9	0.05	-4.3	<0.001
Paternal	0.9	0.8	0.9	0.05	-3.2	0.002	0.7	0.7	0.8	0.05	-6.6	<0.001
Sister	1.4	1.1	1.7	0.11	2.8	0.006	1.1	0.8	1.3	0.12	0.5	0.648
Brother	0.7	0.6	0.8	0.09	-4.2	<0.001	0.8	0.7	1.0	0.08	-2.5	0.013
Nuclear	0.8	0.7	0.9	0.06	-4.3	<0.001	0.5	0.5	0.6	0.07	-9.1	<0.001
Mat orph (bin)	1.2	1.0	1.5	0.1	1.8	0.075	1.1	0.9	1.3	0.09	1.2	0.229
Pat orph (bin)	0.9	0.8	1.0	0.07	-1.5	0.133	0.9	0.8	1.0	0.06	-2.1	0.039
Age (cont.)	1.2	1.2	1.2	0.01	25.9	<0.001	1.1	1.1	1.1	0.01	16.9	<0.001
Calendar year (categorical)												
2004-5	0.6	0.5	0.6	0.07	-8.6	<0.001	0.5	0.5	0.6	0.07	-8.7	<0.001
2006-7	0.7	0.6	0.8	0.06	-5.7	<0.001	0.7	0.6	0.8	0.06	-5.6	<0.001
2008-9	Reference						Reference					
2010-11	0.8	0.7	1.0	0.06	-2.8	0.005	0.8	0.7	0.9	0.06	-2.8	0.005
2012-13	0.8	0.7	0.9	0.06	-2.9	0.003	0.7	0.7	0.8	0.06	-4.8	<0.001
2014-15	0.6	0.6	0.7	0.06	-7.2	<0.001	0.6	0.5	0.7	0.07	-8.4	<0.001
2016-17	0.5	0.4	0.6	0.07	-10.8	<0.001	0.5	0.4	0.5	0.07	-11	<0.001
Distance to tarmac road in km (cont. int.)	1.0	1.0	1.0	0.01	0.7	0.472	1.0	1.0	1.0	0.01	0.2	0.809

	Children (long independent moves)											
	Female (3425 moves)						Male (3321 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Population density within 250m (categorical)												
<50p	1.1	1.0	1.2	0.05	2.2	0.025	1.1	1.0	1.3	0.05	2.6	0.009
50-149p	Reference						Reference					
150-299p	1.1	1.0	1.2	0.05	1.2	0.238	1.2	1.1	1.3	0.05	3.0	0.003
>=300p	1.1	1.0	1.2	0.06	1.9	0.059	1.1	1	1.2	0.06	1.5	0.134
Father secondary education (categorical)												
None	Reference						Reference					
Secondary	1.0	1.0	1.1	0.04	1.0	0.339	0.9	0.8	1.0	0.05	-1.9	0.052
Unknown	3.5	2.9	4.3	0.1	12.5	<0.001	3.3	2.7	4.0	0.1	11.5	<0.001
Mother secondary education (categorical)												
None	Reference						Reference					
Secondary	1.0	0.9	1.1	0.05	-0.9	0.387	1.0	0.9	1.1	0.05	0.7	0.5
Unknown	1.0	0.8	1.2	0.12	-0.3	0.745	1.1	0.9	1.4	0.12	0.8	0.399
Household head employment score (categorical)												
Low	1.1	1.0	1.3	0.08	1.6	0.108	1.2	1.0	1.4	0.08	2.2	0.027
Medium	Reference						Reference					
High	1.0	0.9	1.1	0.05	-0.9	0.346	1.0	0.9	1.1	0.05	0.5	0.591
Unknown	1.2	0.9	1.6	0.15	1.3	0.202	1.3	1.0	1.7	0.14	1.9	0.055

NB: Models run separately for female children (aged <12 [19,650 individuals]) and male children (aged <16 [21,267 individuals]); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id; short moves are under 4 kilometres; independent moves are without a parent of any age or an adult aged 18 or over

Table 8.5a. Results for short independent outcome for adolescents, from multi-level multinomial regression models

	Adolescents (short independent moves)											
	Female (4458 moves)						Male (2492 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Household composition												
Par & siblings	Reference						Reference					
Sister's family	1.3	1.1	1.5	0.08	3.7	<0.001	1.0	0.9	1.2	0.09	0.4	0.682
Brother's family	1.9	1.6	2.3	0.10	6.4	<0.001	1.7	1.4	2.1	0.11	5.2	<0.001
Mother & siblings	1.4	1.2	1.6	0.06	5.4	<0.001	1.0	0.9	1.2	0.08	0.3	0.756
Fath & stepmoth	2.0	1.7	2.3	0.08	8.8	<0.001	1.6	1.4	1.9	0.09	5.7	<0.001
Maternal	1.9	1.6	2.1	0.07	8.3	<0.001	1.6	1.3	1.9	0.09	5.2	<0.001
Paternal	2.7	2.3	3.1	0.08	12.2	<0.001	1.7	1.4	2.1	0.10	5.5	<0.001
Spouse	0.3	0.3	0.4	0.07	-14.8	<0.001	0.2	0.2	0.3	0.12	-13.4	<0.001
Other	1.7	1.4	1.9	0.08	6.0	<0.001	1.1	0.9	1.3	0.10	0.5	0.590
Own child (bin)	0.8	0.7	0.9	0.08	-2.8	0.005	0.4	0.3	0.5	0.12	-8.1	<0.001
Presence in household (all individual continuous variables)												
Under 1 year	1.0	1.0	1.1	0.04	1.1	0.260	1.1	1.0	1.3	0.06	1.9	0.059
1-4 years	1.0	1.0	1.1	0.02	1.6	0.103	1.0	0.9	1.1	0.03	-0.2	0.857
5-18 years	1.0	1.0	1.0	0.01	-0.1	0.894	1.1	1.0	1.1	0.01	4.9	<0.001
19-59 years	0.9	0.9	1.0	0.02	-3.1	0.002	0.8	0.8	0.9	0.02	-7.9	<0.001
60 years & over	0.9	0.8	1.0	0.04	-2.8	0.005	0.9	0.8	1.0	0.04	-2.1	0.032
Presence of family type within 250m (all individual binary variables)												
Maternal	1.2	1.1	1.3	0.05	3.3	0.001	1.1	1.0	1.2	0.06	1.2	0.235
Paternal	1.0	0.9	1.1	0.04	0.8	0.445	1.0	0.9	1.1	0.05	-0.4	0.683
Sister	1.1	1.0	1.3	0.06	2.2	0.028	0.9	0.8	1.1	0.06	-1.1	0.274
Brother	1.2	1.0	1.3	0.05	2.9	0.004	1.1	1.0	1.2	0.05	1.2	0.248
Nuclear	1.1	1.0	1.2	0.06	1.6	0.104	1.0	0.9	1.2	0.06	0.6	0.531
Mat orphan (bin)	1.2	1.0	1.4	0.08	2.5	0.012	1.1	0.9	1.3	0.08	0.8	0.440
Pat orphan (bin)	1.0	0.9	1.1	0.05	-0.7	0.483	1.0	0.9	1.1	0.06	0.3	0.768
Age (continuous)	1.1	1.1	1.1	0.01	18.5	<0.001	1.1	1.1	1.1	0.01	17.3	<0.001
Calendar year (categorical)												
2004-5	1.1	1.0	1.2	0.06	1.2	0.227	1.3	1.1	1.6	0.08	3.7	<0.001
2006-7	0.9	0.8	1.0	0.06	-2.0	0.044	1.1	1.0	1.3	0.08	1.6	0.099
2008-9	Reference						Reference					
2010-11	0.9	0.8	1.0	0.06	-2.2	0.031	0.8	0.7	1.0	0.09	-2.1	0.033
2012-13	0.8	0.7	0.9	0.06	-3.3	0.001	1.0	0.9	1.2	0.08	0.2	0.833
2014-15	0.7	0.6	0.8	0.06	-6.7	<0.001	1.1	0.9	1.2	0.08	0.6	0.518
2016-17	0.6	0.6	0.7	0.06	-7.9	<0.001	1.0	0.9	1.2	0.08	0.5	0.641

	Adolescents (short independent moves)											
	Female (4458 moves)						Male (2492 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Distance to tarmac road in km (cont. integer)	1.0	1.0	1.0	0.01	-1.4	0.160	1.0	1.0	1.0	0.01	3.9	<0.001
Population density within 250m (categorical)												
<50p	0.9	0.8	1.0	0.05	-2.7	0.006	1.0	0.9	1.1	0.06	0.3	0.795
50-149p	Reference						Reference					
150-299p	1.0	0.9	1.1	0.04	0.0	0.962	1.0	0.9	1.2	0.06	0.7	0.504
>=300p	0.9	0.9	1.0	0.05	-1.1	0.266	1.1	1.0	1.3	0.06	1.7	0.092
Father secondary education (categorical)												
None	Reference						Reference					
Secondary	1.0	0.9	1.1	0.04	0.0	0.975	0.9	0.8	1.0	0.05	-1.9	0.060
Unknown	1.1	0.9	1.4	0.12	0.7	0.488	0.7	0.5	1.0	0.15	-2.2	0.030
Mother secondary education (categorical)												
None	Reference						Reference					
Secondary	1.1	1.0	1.2	0.04	1.8	0.078	0.9	0.9	1.1	0.06	-1.0	0.314
Unknown	0.7	0.5	0.9	0.14	-2.5	0.013	0.7	0.5	1.0	0.17	-1.9	0.062
Household head employment score (categorical)												
Low	1.2	1.1	1.4	0.06	3.1	0.002	0.9	0.8	1.1	0.09	-1.0	0.325
Medium	Reference						Reference					
High	0.8	0.7	0.9	0.05	-4.1	<0.001	0.8	0.7	1.0	0.07	-2.5	0.011
Unknown	1.7	1.3	2.2	0.14	3.7	<0.001	1.1	0.8	1.8	0.22	0.6	0.519

NB: Models run separately for female adolescents (aged 12-24 [13,748 individuals]) and male adolescents (aged 16-28 [10,208 individuals]); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id

Table 8.5b. Results for long independent outcome for adolescents, from multi-level multinomial regression models

	Adolescents (long independent moves)											
	Female (6327 moves)						Male (3217 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Household composition												
Par & siblings	Reference						Reference					
Sister's family	1.6	1.4	1.7	0.06	7.7	<0.001	1.5	1.3	1.8	0.09	4.7	<0.001
Brother's family	1.7	1.4	2.0	0.09	5.9	<0.001	2.5	2.0	3.0	0.1	9.0	<0.001
Mother & siblings	1.3	1.2	1.4	0.05	5.3	<0.001	1.5	1.3	1.8	0.08	5.5	<0.001
Fath & stepmoth	2	1.7	2.2	0.06	11.1	<0.001	2	1.7	2.4	0.08	8.3	<0.001
Maternal	2.1	1.9	2.3	0.06	12.9	<0.001	2	1.7	2.4	0.09	8.3	<0.001
Paternal	2.4	2.2	2.7	0.06	14.3	<0.001	2.4	2.0	2.9	0.09	9.4	<0.001
Spouse	0.3	0.3	0.4	0.06	-17.2	<0.001	0.7	0.6	0.9	0.11	-3.4	0.001
Other	2.2	1.9	2.5	0.07	12.1	<0.001	3.2	2.7	3.8	0.09	13.5	<0.001
Own child (bin)	0.9	0.8	1.0	0.07	-2.2	0.027	1.0	0.8	1.2	0.11	-0.1	0.935
Presence in household (all individual continuous variables)												
Under 1 year	0.9	0.8	0.9	0.04	-4.1	<0.001	1.1	1.0	1.2	0.05	1.1	0.279
1-4 years	1.0	1.0	1.0	0.02	-0.3	0.786	0.9	0.9	1.0	0.03	-3.3	0.001
5-18 years	1.0	1.0	1.0	0.01	-2.8	0.006	1.0	1.0	1.0	0.01	-0.4	0.716
19-59 years	1.0	1.0	1.0	0.01	-1.3	0.203	1.0	1.0	1.0	0.02	-0.4	0.696
60 years & over	0.9	0.9	1.0	0.03	-2.7	0.008	1.0	0.9	1.1	0.04	-0.1	0.898
Presence of family type within 250m (all individual binary variables)												
Maternal	0.9	0.8	1.0	0.04	-2.1	0.032	0.9	0.8	1.0	0.05	-1.5	0.127
Paternal	0.9	0.9	1.0	0.04	-2.4	0.015	0.8	0.7	0.8	0.05	-5.6	<0.001
Sister	1.0	0.9	1.1	0.05	0.5	0.628	1.0	0.9	1.1	0.06	0.1	0.914
Brother	1.0	0.9	1.1	0.04	0.3	0.779	0.8	0.8	0.9	0.05	-3.8	<0.001
Nuclear	0.9	0.8	1.0	0.05	-2.2	0.03	0.7	0.6	0.7	0.07	-6.5	<0.001
Mat orphan (bin)	1.1	1.0	1.2	0.06	1.6	0.103	1.1	0.9	1.2	0.07	0.7	0.487
Pat orphan (bin)	1.0	0.9	1.1	0.04	0.1	0.907	1.0	0.9	1.1	0.05	0.01	0.982
Age (continuous)	1.1	1.1	1.1	0.01	23.8	<0.001	1.1	1.1	1.1	0.01	14.4	<0.001
Calendar year (categorical)												
2004-5	0.7	0.6	0.7	0.05	-8.1	<0.001	0.5	0.4	0.5	0.07	-10.3	<0.001
2006-7	0.8	0.7	0.9	0.05	-4.9	<0.001	0.6	0.6	0.7	0.07	-6.6	<0.001
2008-9	Reference						Reference					
2010-11	0.9	0.8	0.9	0.05	-3.4	0.001	0.9	0.8	1	0.06	-2.2	0.031
2012-13	0.8	0.7	0.9	0.05	-5.1	<0.001	0.7	0.6	0.8	0.07	-4.6	<0.001
2014-15	0.7	0.6	0.7	0.05	-8.1	<0.001	0.8	0.7	0.9	0.06	-4.2	<0.001
2016-17	0.5	0.5	0.6	0.05	-12.8	<0.001	0.5	0.5	0.6	0.07	-9.2	<0.001

	Adolescents (long independent moves)											
	Female (6327 moves)						Male (3217 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Distance to tarmac road in km (cont. integer)	1.0	1.0	1.0	0.01	3.8	<0.001	1.0	1.0	1.0	0.01	-1.0	0.301
Population density within 250m (categorical)												
<50p	1.1	1.1	1.2	0.04	3.3	0.001	1.1	1	1.2	0.05	1.4	0.164
50-149p	Reference						Reference					
150-299p	1	1	1.1	0.04	0.6	0.539	1.1	1	1.3	0.05	2.8	0.006
>=300p	1	0.9	1.1	0.04	-0.5	0.643	1.1	1	1.3	0.06	2.2	0.026
Father secondary education (categorical)												
None	Reference						Reference					
Secondary	0.8	0.8	0.9	0.03	-5.6	<0.001	1	0.9	1.1	0.04	-0.2	0.87
Unknown	2.2	1.9	2.5	0.08	10.3	<0.001	2.6	2.2	3.2	0.09	10.4	<0.001
Mother secondary education (categorical)												
None	Reference						Reference					
Secondary	0.9	0.8	0.9	0.04	-3.6	<0.001	0.9	0.8	1	0.05	-2.6	0.01
Unknown	1.7	1.5	2	0.08	6.6	<0.001	1.9	1.6	2.3	0.1	6.4	<0.001
Household head employment score (categorical)												
Low	1.1	1	1.2	0.06	1.9	0.064	1.2	1	1.3	0.07	2	0.045
Medium	Reference						Reference					
High	0.9	0.8	1	0.04	-2.6	0.01	1.1	1	1.2	0.05	1.6	0.103
Unknown	0.8	0.6	1	0.13	-2.1	0.033	1.1	0.8	1.5	0.15	0.7	0.476

NB: Models run separately for female adolescents (aged 12-24 [13,748 individuals]) and male adolescents (aged 16-28 [10,208 individuals]); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id

Table 8.6a. Results for short accompanied outcome for adolescents, from multi-level multinomial regression models

	Adolescents (short accompanied moves)											
	Female (2039 moves)						Male (1378 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Household composition												
Parents & siblings	Reference						Reference					
Sister's family	0.8	0.6	1.1	0.14	-1.6	0.116	0.8	0.6	1.2	0.17	-1.1	0.285
Brother's family	1.6	1.1	2.4	0.21	2.2	0.027	2.2	1.6	3.2	0.18	4.4	<0.001
Mother & siblings	2.1	1.7	2.5	0.09	7.7	<0.001	1.9	1.5	2.4	0.13	4.9	<0.001
Fath & stepmother	1.5	1.1	1.9	0.13	3.0	0.002	1.8	1.4	2.4	0.15	4.1	<0.001
Maternal	1.0	0.8	1.2	0.12	-0.2	0.836	0.9	0.7	1.3	0.16	-0.5	0.595
Paternal	1.0	0.7	1.4	0.17	-0.1	0.903	1.7	1.2	2.4	0.18	2.9	0.003
Spouse	4.2	3.5	5.2	0.10	14.1	<0.001	5.6	4.2	7.5	0.15	11.9	<0.001
Other	0.6	0.4	1.0	0.21	-2.1	0.040	0.9	0.5	1.4	0.24	-0.6	0.535
Own child (binary)	1.3	1.0	1.8	0.14	2.2	0.028	1.3	1.0	1.6	0.12	2.0	0.041
Presence in household (all individual continuous variables)												
Under 1 year	1.1	1.0	1.3	0.05	2.6	0.010	1.2	1.1	1.4	0.07	3.4	0.001
1-4 years	1.0	0.9	1.1	0.03	-0.1	0.955	1.0	0.9	1.1	0.04	-0.1	0.918
5-18 years	1.0	1.0	1.1	0.02	1.3	0.181	0.9	0.9	1.0	0.02	-2.5	0.013
19-59 years	1.2	1.1	1.3	0.02	7.3	<0.001	1.3	1.3	1.4	0.03	10.9	<0.001
60 years and over	1.1	1.0	1.2	0.06	1.6	0.111	1.2	1.1	1.4	0.07	2.8	0.006
Presence of family type within 250m (all individual binary variables)												
Maternal	1.2	1.0	1.3	0.07	2.1	0.035	1.2	1.1	1.4	0.07	2.8	0.005
Paternal	0.6	0.6	0.7	0.07	-6.1	<0.001	0.7	0.6	0.8	0.07	-5.0	<0.001
Sister	1.1	0.9	1.3	0.11	0.6	0.527	0.7	0.6	0.9	0.08	-3.7	<0.001
Brother	0.6	0.5	0.8	0.10	-4.5	<0.001	0.6	0.6	0.7	0.07	-6.6	<0.001
Nuclear	0.7	0.6	0.9	0.12	-2.6	0.010	0.7	0.6	0.9	0.12	-3.0	0.002
Mat orphan (bin)	1.2	0.9	1.5	0.12	1.3	0.207	0.9	0.7	1.2	0.12	-0.5	0.647
Pat orphan (bin)	0.9	0.8	1.0	0.07	-1.3	0.188	1.0	0.8	1.1	0.08	-0.3	0.797
Age (continuous)	0.9	0.9	1.0	0.01	-5.6	<0.001	1.0	1.0	1.0	0.01	0.1	0.925
Calendar year (categorical)												
2004-5	1.3	1.1	1.5	0.09	2.9	0.003	1.3	1.0	1.6	0.11	2.1	0.036
2006-7	1.0	0.9	1.2	0.09	0.3	0.729	1.2	0.9	1.4	0.11	1.5	0.142
2008-9	Reference						Reference					
2010-11	1.0	0.8	1.2	0.09	0.0	0.975	1.0	0.8	1.3	0.11	0.4	0.667
2012-13	0.9	0.8	1.1	0.09	-0.7	0.469	1.1	0.9	1.3	0.11	0.6	0.564
2014-15	0.9	0.8	1.1	0.09	-1.1	0.264	1.0	0.8	1.2	0.11	-0.3	0.800
2016-17	1.0	0.8	1.2	0.09	-0.1	0.956	1.2	1.0	1.5	0.11	1.6	0.104

	Adolescents (short accompanied moves)											
	Female (2039 moves)						Male (1378 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Distance to tar road in km (cont. integer)	1.0	1.0	1.0	0.01	-3.0	0.002	1.0	1.0	1.0	0.01	-0.7	0.464
<50p	0.8	0.7	1.0	0.08	-2.5	0.012	0.7	0.6	0.9	0.09	-3.3	0.001
50-149p	Reference						Reference					
150-299p	1.2	1.0	1.3	0.07	2.3	0.022	1.1	0.9	1.2	0.08	0.8	0.419
>=300p	1.9	1.6	2.1	0.07	9.5	<0.001	1.6	1.3	1.8	0.08	5.7	<0.001
Father secondary education (categorical)												
None	Reference						Reference					
Secondary	1.0	0.9	1.1	0.06	-0.2	0.874	1.1	1.0	1.3	0.07	1.7	0.098
Unknown	1.2	0.8	1.6	0.16	0.9	0.359	0.8	0.6	1.2	0.19	-0.9	0.380
Mother secondary education (categorical)												
None	Reference						Reference					
Secondary	0.9	0.8	1.0	0.06	-1.2	0.240	1.0	0.9	1.2	0.07	0.2	0.842
Unknown	0.7	0.5	1.0	0.19	-1.9	0.062	0.9	0.6	1.4	0.22	-0.5	0.605
Household head employment score (categorical)												
Low	1.1	0.9	1.4	0.11	1.1	0.250	1.1	0.8	1.3	0.12	0.5	0.645
Medium	Reference						Reference					
High	1.9	1.7	2.1	0.06	10.7	<0.001	1.8	1.5	2.0	0.07	7.8	<0.001
Unknown	2.8	2.0	3.8	0.16	6.3	<0.001	2.1	1.3	3.2	0.23	3.3	0.001

NB: Models run separately for female adolescents (aged 12-24 [13,748 individuals]) and male adolescents (aged 16-28 [10,208 individuals]); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id

Table 8.6b. Results for long accompanied outcome for adolescents, from multi-level multinomial regression models

	Adolescents (long accompanied moves)											
	Female (1892 moves)						Male (1321 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Household composition												
Parents & siblings	Reference						Reference					
Sister's family	1.0	0.8	1.3	0.12	0.3	0.781	0.8	0.6	1.1	0.14	-1.3	0.192
Brother's family	1.2	0.7	1.8	0.22	0.7	0.512	1.2	0.8	1.7	0.18	0.9	0.376
Mother & siblings	1.0	0.8	1.2	0.1	-0.1	0.885	0.8	0.6	1.0	0.13	-1.9	0.059
Father & stepmoth	1.2	0.9	1.5	0.12	1.4	0.162	1.1	0.8	1.4	0.13	0.4	0.712
Maternal	0.6	0.5	0.8	0.13	-3.6	<0.001	0.6	0.4	0.8	0.16	-3.6	<0.001
Paternal	0.6	0.5	0.9	0.16	-2.7	0.007	0.6	0.4	1.0	0.21	-2.1	0.035
Spouse	1.2	1.0	1.5	0.1	1.9	0.058	2.1	1.6	2.8	0.15	5.0	<0.001
Other	0.5	0.4	0.8	0.16	-3.7	<0.001	0.5	0.4	0.8	0.2	-3.1	0.002
Own child (binary)	1.4	1.1	1.9	0.15	2.5	0.014	2.5	1.9	3.3	0.15	6.3	<0.001
Presence in household (all individual continuous variables)												
Under 1 year	1.0	0.9	1.1	0.06	-0.6	0.539	1.0	0.9	1.1	0.07	-0.1	0.944
1-4 years	0.9	0.8	0.9	0.04	-3.5	0.001	0.9	0.8	1.0	0.05	-2.4	0.018
5-18 years	0.9	0.9	0.9	0.02	-5.9	<0.001	0.9	0.9	0.9	0.02	-4.8	<0.001
19-59 years	1.2	1.1	1.2	0.03	5.7	<0.001	1.2	1.1	1.3	0.03	5.8	<0.001
60 years and over	0.9	0.8	1.0	0.07	-1.7	0.082	0.8	0.7	0.9	0.08	-3.1	0.002
Presence of family type within 250m (all individual binary variables)												
Maternal	0.6	0.5	0.8	0.09	-4.9	<0.001	0.8	0.7	1.0	0.08	-2.5	0.013
Paternal	0.5	0.4	0.6	0.08	-9.6	<0.001	0.5	0.5	0.6	0.07	-9.0	<0.001
Sister	0.8	0.6	1.0	0.13	-2	0.046	0.8	0.7	1.0	0.09	-2.0	0.044
Brother	0.4	0.4	0.6	0.11	-7.2	<0.001	0.6	0.5	0.7	0.07	-7.0	<0.001
Nuclear	0.7	0.5	0.9	0.13	-2.8	0.005	0.5	0.4	0.7	0.13	-5.2	<0.001
Mat orphan (bin)	0.9	0.7	1.2	0.13	-0.6	0.581	1.0	0.8	1.3	0.12	-0.1	0.9
Pat orphan (bin)	0.9	0.8	1.1	0.08	-0.7	0.467	0.9	0.8	1.1	0.08	-1.3	0.206
Age (continuous)	1.0	1.0	1.0	0.01	-0.1	0.903	1.0	1.0	1.0	0.01	-0.4	0.721
Calendar year (categorical)												
2004-5	0.8	0.7	1.0	0.09	-2.1	0.033	0.6	0.5	0.8	0.11	-4.5	<0.001
2006-7	0.8	0.7	1.0	0.09	-1.9	0.052	0.8	0.6	1.0	0.1	-2.4	0.017
2008-9	Reference						Reference					
2010-11	0.9	0.7	1.0	0.09	-1.5	0.136	0.8	0.7	1.0	0.1	-1.7	0.095
2012-13	0.9	0.7	1.1	0.09	-1.4	0.175	0.9	0.7	1.1	0.1	-1.3	0.206
2014-15	0.9	0.7	1	0.09	-1.7	0.085	0.8	0.6	1	0.1	-2.3	0.023
2016-17	0.7	0.6	0.8	0.09	-4.5	<0.001	0.5	0.4	0.7	0.11	-5.6	<0.001

	Adolescents (long accompanied moves)											
	Female (1892 moves)						Male (1321 moves)					
	OR	95% CI		SE	Z	p	OR	95% CI		SE	Z	p
Distance to tarmac road in km (cont. integer)	1.0	1.0	1.0	0.01	1.4	0.153	1.0	1.0	1.0	0.01	-0.7	0.503
Population density within 250m (categorical)												
<50p	0.8	0.7	1.0	0.07	-2.3	0.023	0.8	0.7	1.0	0.08	-2.0	0.049
50-149p	Reference						Reference					
150-299p	1.1	0.9	1.2	0.06	0.9	0.374	0.9	0.8	1.0	0.08	-1.5	0.128
>=300p	1.2	1.1	1.4	0.07	3.2	0.001	1.0	0.8	1.2	0.08	-0.2	0.844
Father secondary education (categorical)												
None	Reference						Reference					
Secondary	0.9	0.8	1.0	0.06	-2.7	0.006	0.9	0.8	1.1	0.07	-1.0	0.306
Unknown	2.3	1.8	3.0	0.13	6.2	<0.001	2.3	1.7	3	0.15	5.5	<0.001
Mother secondary education (categorical)												
None	Reference						Reference					
Secondary	0.8	0.7	0.9	0.06	-4.4	<0.001	0.9	0.8	1.1	0.08	-0.9	0.373
Unknown	1.2	0.9	1.7	0.14	1.6	0.116	1.8	1.3	2.5	0.16	3.9	<0.001
Household head employment score (categorical)												
Low	1.0	0.8	1.3	0.13	0.02	0.994	1.3	1.0	1.7	0.13	2.0	0.05
Medium	Reference						Reference					
High	2.4	2.2	2.7	0.06	15.6	<0.001	2.5	2.2	2.9	0.07	13.2	<0.001
Unknown	5.7	4.5	7.2	0.12	14.6	<0.001	4.8	3.6	6.5	0.15	10.5	<0.001

NB: Models run separately for female adolescents (aged 12-24 [13,748 individuals]) and male adolescents (aged 16-28 [10,208 individuals]); outcome is move type with ‘no move’ as baseline; models allow for clustering by unique household id and unique participant id

8.5. Discussion

8.5.1. Summary of findings

This analysis on mobility in rural Malawi shows two key periods of mobility in very young childhood and adolescence/young adulthood which relate to leaving home and marriage as part of the transition to adulthood, and young couples moving to establish themselves in the community while their children are young. The tradition of patrilocality can be seen in these data, with young women moving to join their spouses, however the culture of children 'belonging' to the paternal family is less clear with children seeming to be treated differently according to their sex. Moving seems to be strongly linked to local presence of family, in that the presence of family living nearby but not in the same household tends to be associated with lower chances of moving. Mobility is also related to socio-economic status and appears to become less common over time.

8.5.2. Moving is common for young families

The high level of movement in young families can be seen in the high rates of accompanied move for infants and young children, which sharply, and then more gradually, decrease until about the age of 12. High move rates in young children have been found elsewhere (Ford and Hosegood, 2005; Grieger et al., 2013). There is also an increase in accompanied moves in young adulthood. Evidence from the regression modelling shows that adolescents living in a '*Spouse'* household type are more likely to move accompanied (*i.e.* with their spouse) and that presence of very young children was associated with short accompanied moves but not long ones, implying that parents with young families are likely to move around locally but not to migrate long distances. Also, not living with a parent (*i.e.*, with *paternal* household) tended to be associated with lower chance of move than living in a *parents & siblings* household. The descriptive analysis shows that children moving accompanied often do not experience a change in household type implying it is a relocation of the household, rather than it breaking up, although there is also evidence of accompanied moves due to marital dissolution (*i.e.* children commonly move from '*Parents and siblings*' households to '*Mother and siblings*'). A limitation of this dataset is that it is known that men who marry and start a family while still living with their parents are likely to declare themselves a distinct household so, if/when they did move away from the grandparents the move from *paternal* households to *parents and siblings* would not be observed in this data.

8.5.3. Adolescents move as part of transition to adulthood

Both short and long independent move risks rise during adolescence: this rise happens earlier than for accompanied moves and is also earlier for females than males. This is expected given that it is known that women tend to transition to adulthood earlier in this population (McLean et al., 2021b), a pattern which has been found elsewhere (Beegle and Poulin, 2013; Ford and Hosegood, 2005; Grieger et al., 2013). The patrilocal tradition can be observed as female adolescents are much more likely to leave the spousal home to return to the home of another relative (or leave the area), presumably following separation or divorce. Following a breakdown in a marriage, it appears that the men are likely to stay put. Both males and females had a higher chance of independent move when living in household types other than '*parents and siblings'* (excluding *spouse*-type household): a study from another area of Malawi also found that adolescents living with parents were less likely to move than those with a more distant relationship to the household head (Beegle and Poulin, 2013).

Presence of infants aged under one year was associated with lower chance of long independent move for female adolescents, which might be expected if they were 'held back' from their own marriage to care for other young children: for males this effect was also found for presence of one to four-year-old children. It may be that adolescents of both sexes delay long-distance moves to support young children in their households.

8.5.4. Sex differences and cultural aspects

The most obvious difference between male and female children in these data is the higher rates of long independent moves for girls compared to boys from the age of four. We have found previously that there are sex differences in which relatives children and adolescents live with, if not with both parents: boys are more likely to live in 'male' relative household (*i.e.* with brother, father without mother, or paternal relatives) while girls may be more likely to live with a sister (McLean et al., 2021a). This was confirmed in this analysis of mobility where some household moves were common for each sex, *i.e.,* girls were more likely to move to a *maternal* household: as the mother's family is more likely to be further away (due to the patrilocal traditions) this may explain some of the difference in risk of long independent moves, however unfortunately we do not know what type of household most long-distance movers go to. These girls may also be fostered out to households better able to care for them: this is relatively common in sub-Saharan Africa and has been found to be slightly more likely for girls than boys in other settings (Hedges et al., 2019), or maybe being sent to help out in other households either doing house work or caring duties, which also tends to be

more likely for girls (Robson, 2000). Presence of small children was associated with higher chance of long independent move for female children, but not for males, indicated that the female children might be being moved out to make room for younger ones: a study in Malawi found that children were more likely to be fostered out if the mother had just had a baby, but that this depended on the mother's marital status (Grant and Yeatman, 2014).

Children of both sexes have a much higher rate of independent move in certain household types, generally when the mother was not present, which has also been observed elsewhere (Madhavan et al., 2012). It was also found that girls moving accompanied were more likely than boys to be moving with their mother: this may be because when a marriage breaks down and the woman moves out, she may be more likely to take the female children rather than the male.

In general, presence of relatives nearby is associated with lower chance of moving, indicating that family is a strong force in decision-making in this area. Although in this area children traditionally 'belong' to their paternal family, these data clearly show that the maternal family plays an important role. There were more moves to *maternal* households than *paternal* ones, and presence of maternal family nearby but not in the household is associated with higher chance of short move but lower chance of long move indicating that presence of maternal family 'keeps' families living in the same area. Living in a '*Mother & siblings'* household is associated with higher chance of short and long accompanied move, and the descriptive analysis shows that the short moves at least tend to be relocations (*i.e.*, remaining '*Mother & siblings'*) or could be the mother and children moving to live with her family (*i.e.*, *maternal* household), so it may be likely that a proportion of the long moves will be due to the mother returning to her home village. In contrast living in '*Father and step-mother'* household (which will also include single father) was only associated with higher chance of short accompanied move and most of these are relocations (*i.e.*, remaining '*Father & step-mother* type), although it is likely that a man moving back to his parent's property would maintain that he is a separate household so we wouldn't capture this change in this data, it seems likely that if fathers do move following widowhood or divorce it is to elsewhere in the area.

8.5.5. Other factors relating to mobility

Although family is clearly important in decision making about mobility, our analyses also found independent effects of other factors. Moving rates appear to be decreasing since 2008. There could be many reasons for this. It is known that transition to adulthood is becoming later for both sexes with adolescents more likely to stay in school longer before

leaving home and marrying (McLean et al., 2021b), which would affect rates of mobility of young people. This area was also heavily affected by HIV/AIDS but increasing access to ART since the mid-2000s has reduced mortality rates (Price et al., 2017) which will have kept more households intact so children and adolescents have not had to move. Equally, fertility rates have dropped in the area (McLean et al., 2017) so if children and adolescents are moving to free up space in the household this will be required less. There does not appear to be a trend of young people increasingly moving to cities, as might be expected given increasing urbanisation and access to internet/media which has been found in other settings (Hertrich and Lesclingand, 2012). Further analysis once more data become available are needed to assess whether the decrease is a lasting trend.

Household head socio-economic position seems to have little effect on independent move, but those in the most socio-economically advantaged group are most likely to move accompanied. This may be because those in the medium socio-economic group will tend to be farmers whose work is linked to their land, so it would be harder to move, while those in the most advantaged group may move due to work. A study in another area of Malawi also found that adolescents in the wealthiest households were most likely to move (Beegle and Poulin, 2013), and in urban South Africa moving was more likely for children living at the top or bottom of the socio-economic scale (Ginsburg et al., 2009).

8.5.6. Strengths and limitations

This analysis benefitted from a very detailed dataset capturing short and long distance moves with the ability to differentiate between independent and accompanied moves. However, the census is carried out annually, so short-term migrations occurring between the census rounds may be missed. Moves may be wrongly classified as independent or accompanied if move dates of other household members are wrongly captured (though collapsing the data into quarters reduces the reliance on the exact dates being the same) or, for example, if a person is joining a household member who moved shortly before. In addition, external moves may be more likely to be erroneously classed as accompanied as they only need to have reported moving from/to the same town/city rather than to the exact same house. GPS data for households within the area allow for calculations of distances moved and the ability to class people as living near certain relatives. Outside of the HDSS the move distances are estimated, however this should not have resulted in too many misclassifications into short/long as almost all external moves would be classed as long.

Short-term temporary moves or holidays would not be captured in our data: a study in Zambia found that a lot of children spent long periods of their school holidays staying with

relatives, often helping out with chores. They speculate that for some of these children these ‘holidays’ may be the result of a balancing act to allow the children to support the extended family but to keep them in school, whereas otherwise they might be fostered out more permanently (Hunleth et al., 2015).

Family links allow for very detailed analysis of the effects of presence of certain types of family, both in the household and nearby. However, family links are not available for all participants; this is most likely for people who have recently moved into the area, *i.e.* women who have moved into the area for marriage will be less likely to appear in the regression model due to being less likely to have at least one parent ID. Even if the individuals have parental IDs available, all their relatives may not, meaning that they will appear to have fewer relatives living nearby. Participants who have moved in will be less likely to have parental IDs, and if a previous move is related to a later move this may cause bias in the estimations. In our analysis, we assume that family members nearby will be in contact and will give and receive support in some manner which may result in them influencing migration decisions, however this may not be the case, and people without family may receive the same support from non-related people which might attenuate any effects of the presence of family.

Further factors beyond what was examined in the model might have an effect but were not available for all participants over the whole time period: missing data from surveys is likely to be associated with the outcome as more mobile households/individuals are more likely to be missed (as was observed in the household socio-economic status variable used in the model). In-depth interpretation of all the factors included in the model was also beyond the scope of this paper. This analysis used only data from single time snapshots just before the move or just after. As most moving decisions would be made over a longer period, examining exposures over a longer period before the move may provide greater understanding, however due to the nature of HDSS data this would not be possible.

8.5.7. Conclusion

Using detailed longitudinal data from a rural HDSS in northern Malawi, we have shown that mobility is very common among young people. While some of these moves are clearly household relocations, children not uncommonly move independently of their parents, with 20% of moves involving unaccompanied children. Overall, we find considerable complexity in movement patterns, though some trends emerge. For example, sex differences in mobility are noticeable, with girls being more likely to move than boys from the age of four; and, in terms of age patterns, we replicate previous findings of high mobility for very young children

and adolescents. But we also find complexity in the households which young people move to and from, and that the maternal family is clearly important in this traditionally patrilocal community. We further find that moves become less common over calendar time, and that we don't observe an increasing trend for young people to move to more urban areas, as might be expected from a context of rapid urbanisation, however it has previously be shown that increasing urbanisation can be mostly explained by natural population increases in urban areas and reclassification of rural areas, rather than rural-urban migration (Farrell, 2017).

Competing interests

No competing interests were disclosed.

Grant information

This work is supported by The Wellcome Trust (098610; 217073; through funds awarded to Amelia Crampin and The Karonga HDSS).

Data availability statement

Underlying data

Due to the nature of the dataset (containing exact GPS coordinates of individuals households and potentially unique patterns of local relatives), it would not be possible to anonymise it in such a way that would sufficiently protect the participants' privacy and allow for useful analyses. MEIRU are, however, keen to share data and collaborate with bona fide researchers and students at universities and research institutes. Interested parties should contact the first author [EM] through info@meiru.mw in the first instance, quoting the paper title. After a discussion of data needed, a signed data transfer agreement would be required. Metadata on the MEIRU dataset which formed the basis of these analyses can be found, along with information on other studies, on MEIRU's data catalogue (http://kpsmw.lshtm.ac.uk/nada/index.php/catalog/13),

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	145184	**Title**	Ms
First Name(s)	Estelle		
Surname/Family Name	McLean		
Thesis Title	Demonstrating the value of Health and Demographic Surveillance Site data for complex secondary analyses, illustrated with analyses of young people's living arrangements and transitions to adulthood.		
Primary Supervisor	Rebecca Sear		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Not yet decided
Please list the paper's authors in the intended authorship order:	Estelle McLean, Albert Dube, Emma Slaymaker, Maria Sironi, Amelia Crampin, Rebecca Sear

Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I developed the concept, carried out all data manipulation and analyses, drafted the manuscript and made revisions following comments from co-authors

SECTION E

Student Signature	
Date	27 July 2023

Supervisor Signature	
Date	27 July 2023

9. Divorce and the transition to adulthood in rural Malawi

Estelle McLean[1,2], Albert Dube[2], Emma Slaymaker[1], Maria Sironi[3], Amelia Crampin[1,2], Rebecca Sear[1]

1. Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine
2. Malawi Epidemiology and Intervention Research Unit
3. University College London

9.1. Abstract

Marriage is almost universal in rural Malawi and tends to happen young. It is one of the key markers of adulthood for both men and women. Divorce is also very common, and some research has suggested that early divorce may 'reset' the transition to adulthood, allowing young people to return to school, especially men, as young divorced women may be disadvantaged by child-care responsibilities. We use longitudinal data from the Karonga Health and Demographic Surveillance Site in rural Northern Malawi to investigate how gender and child-bearing are associated with divorce at a young age (under 18 for women and 22 for men) and its impact on markers of transitions to adulthood. Rates of divorce were higher for women than men, but men were more likely than women to remarry following divorce. Compared to their never married contemporaries, divorced men were more likely to marry, less likely to live with family, and less likely to attend school. Divorced women were as likely to live with family and to marry compared to never married women of the same age; divorced women, however, were less likely to attend school. Having children was associated with increased likelihood of divorce for both men and women, however for men there was no evidence of associations between having children and subsequent outcomes, while for women having children was associated with lower chances of remarriage and attending school. A divorce at a young age did not appear to alter the transition to adulthood, especially for men. While divorced women appeared to have some degree of 'reset' as they returned to live with family, there is not evidence that it was a true 'reset' as they did not return to school.

9.2 Introduction

In rural Malawi, marriage is expected and almost universal, often happening at a relatively young age, especially for women: the 2015-2016 Malawi Demographic and Health Survey

reported that nearly half of Malawian women were married by age 18 (National Statistical Office, 2016). Young people tend to choose their own partners, however a marriage might be expected by parents in the event of a pregnancy (Ansell et al., 2018; Melnikas et al., 2022). Pre-marital sex is common and increasing: research in the same setting as this analysis found that younger people were more likely to report a boy or girlfriend as their first partner than a spouse (Glynn et al., 2010). Marriage is one of the main traditional markers of moving from adolescence to adulthood, and many young people enjoy the independence and responsibilities it grants them; though some do report feeling a burden especially if the marriage was due to pregnancy (Ansell et al., 2018).

Divorce is also common in Malawi at all ages (Clark and Brauner-Otto, 2015), but often occurring within the first 3 years of the union (Bertrand-Dansereau and Clark, 2016; Malinga John, 2022) particularly if the couple are young, have not known each other long or if the marriage was instigated or rushed into due to a pregnancy (Bertrand-Dansereau and Clark, 2016). Pre-marital conception or child-bearing has been found to be associated with divorce in young people (Odimegwu et al., 2017; Smith-Greenaway et al., 2021). Some research has suggested that divorce at a young age may ‘reset’ the transition to adulthood for some young people in Malawi, allowing them to make different life choices (Grant and Pike, 2019). As marriage and pregnancy are common reasons given for school drop-out, an early divorce may enable the young person to restart school and improve their prospects. Indeed, since the legal marriage age was raised to 18 in 2017 (Daniel, 2017), some communities in Malawi have been annulling marriages involving girls aged under 18 and encouraging them to return home with the aim of them returning to school (Melnikas et al., 2021). However, it has also been found in Tanzania that lack of education prospects may encourage marriage, rather than the other way around, so divorce may not result in a return to school (Stark, 2018). Returning to the natal home may therefore represent an unhelpful regression for young divorcees: a qualitative study of young people in Zambia found that, while divorce provided relief of escaping a bad relationship, many experienced an undesired reduction in independence following having to return to the parental home (Mweeba and Mann, 2020). It has also been suggested that divorce may be used as a strategic tool, particularly for young women, to gain independence and empowerment: the first marriage being the first stepping stone away from parental control then the divorce to avoid spousal control (Reniers, 2003); this has also been noted in Burkina Faso (Guirkinger et al., 2021). Equally some young people may not have the option of returning to their parents and may struggle with that independence (Grant and Pike, 2019).

In Malawi, divorce tends to be accessible for both men and women and is not heavily stigmatised: anecdotally young people might experience more stigma for getting married early than for getting divorced (personal communications with field-workers living and working within the community from which the data for this analysis is drawn). Though in other areas of Malawi there were reports of negative perceptions in the communities of girls who had been removed from marriages (Melnikas et al., 2021). Couples of any age may be advised and counselled by family and the church elders to work through any marital issues, however after a period of time it is culturally acceptable to part. Young children are expected to stay with the mother following a divorce, potentially leaving her at a disadvantage. The ex-husband is often seen as comparatively unburdened, able to return to school and/or remarry without consequence (same personal communications as above). The lack of consequences of divorce for men has also been suggested in some published research, for example a review of data from some Africa societies which use ‘lobola’ or bride-price found that divorced women tended to face stigma, while men could divorce and remarry with little or no prejudice (Kgadima and Leburu, 2022). Research on whether divorce is associated with subsequent outcomes for young men in Africa is sparse, but research on young fatherhood has shown, in contrast to the above perspectives, there can be impact on education and remarriage: in Malawi continued schooling was discouraged by family and community members for boys once they become fathers (Parrot et al., 2015); in South Africa young fathers experienced stigma and performed less well academically (Mukuna, 2020); and negative outcomes of early fatherhood such as school drop-out and health outcomes were also found for young men in a study in Ethiopia, India, Peru & Vietnam (Jeong, 2021).

9.3. Objectives

In this analysis we use longitudinal data on young men and women from the Karonga Health and Demographic Surveillance Site (HDSS) in rural Northern Malawi to investigate the associations between gender and child-bearing on divorce at a young age and divorce’s impact on markers of transitions to adulthood. While this data resource was not set up to answer this specific question, it provides some unique aspects that make it useful for studying such topics. First, continuous prospective follow-up of participants improves the likelihood of capturing brief marriages. Second known spouse links and precise information on when household members move in and out of households can not only identify marital breakdown, and differentiate between reconciliation following separation or divorce, and remarriage to another person.

9.4 Literature review

In addition to gender and child-bearing, previous research on divorce (at all ages) in sub-Saharan Africa has identified other predictors of divorce which fall into 3 rough, overlapping categories: socio-economic position, factors relating to the marriage and familial/kinship factors.

9.4.1. Socio-economic position

Urbanisation and modernisation has long been associated with marriage instability and divorce (Goode, 1993), though the direction of effects may be variable. Increasing access to education and employment may lead to an increase in divorce if it gives people the confidence and opportunity to leave unsatisfactory marriages, however the same factors could potentially lead to a decrease in divorce if the increased empowerment allows people to choose more compatible partners and/or gives them improved tools and resources to navigate relationships. Indeed, in a country level examination of predictors of divorce rates, Clark et al found that higher levels of women's employment was associated with higher levels of divorce, but higher levels of women's education was associated with lower divorce rates (Clark and Brauner-Otto, 2015). However, another study in Malawi found no association between education level and divorce (Spell et al., 2012). Low socio-economic position of the marital home has been found to be associated with greater chance of divorce (Porter et al., 2004) and lack of resources was given as a reason for young women to leave, or be removed from a marriage in qualitative work in Zambia (Mweeba and Mann, 2020).

9.4.2. Marital factors

Age at marriage has been shown to be a key risk factor, with younger age at first marriage consistently found to be associated with higher divorce risk (Clark and Brauner-Otto, 2015; Reniers, 2003; Tilson and Larsen, 2000). Polygyny was found to be a risk factor for divorce for women in Malawi (Reniers, 2003), but not in Ghana (Takyi, 2001), and large age difference between the spouses was found to be associated with less chance of divorce in Malawi (Reniers, 2003). A Malawian study found that marriages where partners had not known each other for a long time were more likely to end, potentially due to both lack of family involvement and support, and emotional bond between the pair (Bertrand-Dansereau and Clark, 2016). The Zambian qualitative study suggested that the reasons for unhappiness in marriages between young people were similar to those in older people (infidelity, lack of money/support, violence etc.) but because they are young they lack the resources and

experiences to deal with them (Mweeba and Mann, 2020). In countries affected by the HIV pandemic, marriage may be seen as a risk factor, but also as a safe haven from the virus. Actual HIV infection (Porter et al., 2004) and perceived risk of future HIV infection (Grant and Soler-Hampejsek, 2014) have been found to be predictors of divorce.

9.4.3. Familia/kinship factors

In Ghana and Malawi it has been found that divorce is more likely for couples from a matrilineal background, or if the couple are living with or near the wife's family (Reniers, 2003; Takyi, 2001). It has been suggested that women are more empowered in the relationship by being near to her family. In patrilineal areas children traditionally 'belong' to the paternal family (Mwambene, 2012), while in matrilineal areas the opposite is true so a woman may divorce with less fear of losing her children. In patrilineal areas of Malawi (including the area where the present study is set) it is traditional for the groom to pay 'lobola' or bride price to the bride's family which may be expected to be paid back if the marriage fails, leading some families reluctant to support divorce if it had been paid (Bertrand-Dansereau and Clark, 2016). Regardless of the kinship system, a qualitative study found that the young women choosing to leave their marriage always had support from their natal family to move back home, or even were removed from a violent or bad marriage by a member of their family. This same study also noted that couples who divorced tended to be those who had been living independently as a couple/family (Mweeba and Mann, 2020), rather than still attached to one of their families, implying that as well as family factors that assist to dissolve the marriage, other family factors may help to keep the couple together. In Malawi marriages are traditionally negotiated by senior relatives ('ankhoswe') (Chimango, 1977), who may also act as mentors during the marriage. Anecdotally, elopement (marriage without these negotiations or traditional ceremony) is becoming more common, and lack of involvement of ankhoswe has been shown to be associated with higher chance of divorce (Bertrand-Dansereau and Clark, 2016).

9.5 Methods

9.5.1. Context

The Karonga Health and Demographic Surveillance Site (HDSS) was established in 2002 in the southern part of the Karonga district in northern Malawi (Crampin et al. 2012). The area is largely rural with one semi-urban trading town, several smaller market villages and one port on Lake Malawi. The majority of the population engage in subsistence farming or fishing. The main ethnic group are Tumbuka, who have followed patrilineal and patrilocal

custom since the 19th century: women tend to move to their husband's village when they marry (Malawi Human Rights Commission 2006). In the event of divorce or even paternal death, children considered to be old enough to be away from their mother may be required to live with their father's family (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.

The HDSS covers an area of 150km^2 and by 2016 had over 40,000 people under surveillance, with very high response rates. Household membership is defined by the participants with guidance from trained fieldworkers: all household members must usually live in the dwelling/compound together and recognise the same household head. Men with more than one wife who do not live in the same location are assigned to be living in each wife's household; all other individuals may only belong to one household. Births and deaths are captured monthly through a system of local 'key informants', while migrations are captured annually through visits to all households. If a whole household moves, then this information is gathered from the key informants. When a new household member 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 (even if they are not currently HDSS participants). 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. This information was used to identify all family links (by blood and by marriage) between all HDSS participants. GPS coordinates are recorded for each household when they are registered and if they move: this allows for generation of variables to indicate the presence of certain relatives registered as living near each index at any point in time. Annual surveys capture individual and household socio-demographic indicators including schooling, occupation and marital status.

9.5.2. Ethics

Household heads provide written informed consent on behalf of the whole household to participate in the Karonga HDSS, which may be rescinded at any time for any reason. The HDSS is regularly reviewed and approved by the Malawian National Health Sciences Review Committee (approval #419), and the London School of Hygiene and Tropical Medicine Ethics Committee (approval #5081).

9.5.3. Datasets

Data on HDSS participants are gathered as event reports (births, deaths and migrations) and surveys. The event data is used to create continuous episodes, and the survey data assumed to be valid for dates within certain periods before and/or after the survey date (periods depend on the type of data). Due to the complexity of the exposure data used in these analyses, the episodic data were reduced to a panel dataset of one data point per quarter (15th of middle month of each quarter) per person. Marital status was assigned to each quarter based on data from annual surveys (which may have been a self- or proxy-report), from dates of marriage reported in the same surveys and from whether they were living with their reported spouse. Precise dates of marriage and divorce are not available so the above data sources were used to assign marital status as follows: people were assigned as ‘married’ if they were reported as such and/or were living with their spouse, and they were assigned as ‘divorced’ if they were reported as such or were not living with their spouse. Marriage or divorce are common reasons for migration in and out of the area, especially for women. To avoid missing events the reason for migration was also used to identify marriage and divorce events. The reason for moving may not always relate to the index person (i.e. a young woman moving because her mother divorced would have ‘divorce’ as the reason for moving) and therefore, only independent moves with reason of marriage or divorce were included (McLean et al., 2023). Variables related to having children were generated from dates of birth of all HDSS members and their links to parent IDs. This means that only parents of children registered in the HDSS are recognised. Also available for each individual for each quarter are household composition (McLean et al., 2021a), current schooling status and highest schooling achieved, parental education, household occupation (a composite variable taking into account the reported occupation of all household members), and distance to the main road. Data from 2004-2017 is included. The markers of adulthood examined in this analysis were marriage, living independently (i.e. with self or spouse as head of household rather than a relative’s household) and attending school. These markers were chosen due to data availability and as they are important transitions in this community. Other potential markers, for example working and taking on community roles were not available.

9.5.4. Analyses

All data manipulation and analysis used Stata 16.1. All analyses described below were carried out separated for men and women. As the HDSS is an open cohort, participants may move in and out of the area at any time. This means that some participants will have more complete data than others, some participants will have all their marriage/divorce transitions

observed and other will not. Including only participants with full data would introduce bias, as they are likely to different from those who have moved in or out. For this reason, 4 samples of the data are used to examine each transition, these are described below. The focus of the analyses was on marriages and divorces occurring at a young age so relatively young age cut-offs were used, however it was also necessary to consider the data available. Age cut-offs of 18 for women and 22 for men were used which allowed for large enough samples but still selected those experiencing the transitions at a 'young' age.

Rates and predictors of divorce at young age (sample 1)

We defined early marriages as those that occurred before the age of 18 for women and 22 for men. The ages are different as it is known that women tend to marry earlier than men in the area (McLean et al., 2021b). Participants whose first marriages could be identified (transition from 'never married' to 'married' observed in the dataset), were included in a longitudinal dataset which started at the beginning of the first marriage and ended with divorce, out-migration, death or end of analysis (maximum 3 years). The 3-year time period was chosen as previous literature showed this period to be when most divorces happen (Bertrand-Dansereau and Clark, 2016; Malinga John, 2022), and the focus of the analysis was on divorce at a young age. The main explanatory variable was a time-varying variable indicating timing of the first birth with the categories of 'None' (baseline), 'Pre-marital conception' (first child born before or in the first 9 months of the marriage) and 'Conception in marriage' (first child born after first 9 months but before end of marriage or analysis). Other explanatory variables included age at marriage; calendar year; schooling status: left without completing primary (baseline), currently attending (either primary or secondary), completed primary, and left with some/complete secondary; whether mother or father attended any secondary school; composite variable of reported occupation of household members: only farming (baseline), any irregular wage earner, any regular wage earner and none or unknown; and living within 1km of the main tarmac road; living with only 1 or no parents before marriage [the 2 categories were combined to simplify the model as the effects were the same]; and living arrangements during the marriage: no other family apart from spouse/children (baseline), with or very near own family and with or very near spouse's family; spouse age difference: 5 or more years younger (men only), similar age (baseline), and 5 or more years older (women only); whether spouse was previously married, and if the marriage was polygynous (women only as male first marriages very unlikely to be polygynous). For each of the main exposure categories crude divorce rates were estimated and compared and Kaplan-Meier plots run by child status. As the Kaplan Meier plots indicated different effects of child status by year of marriage, piecewise exponential

regression models were fitted firstly with child status and year of marriage separately, and then including an interaction between year of marriage and child variable.

Associations between divorce at young age and markers of transitions to adulthood: remarriage (sample 2 & 3)

Participants who were reported to be divorced before the age of 18 for women and 22 for men were included in a longitudinal dataset which started at the first report of divorce and ended with remarriage, out-migration, death or end of analysis (3 years following first report of divorce) (sample 2). The main explanatory variable was a time-varying variable indicating number of children with the categories of 'None' (baseline), 'at least 1 child', and 'expecting a child' (period 6 months before the birth of a child). The 'expecting' category was defined as 6 months before a birth rather than 9 months as the pregnancy might not be acknowledged in the first trimester. Crude remarriage rates and Kaplan-Meier plots were examined for each category, and piecewise exponential regression models run, controlling for age, year, own schooling, parent's schooling, household occupation and living with parents. For comparison, identical analyses were carried out on a similar dataset including an age/sex frequency matched sample of young people who were never married: for these analyses the outcome was first marriage (sample 3).

Associations between divorce at young age and markers of transitions to adulthood: living arrangements and schooling (sample 4)

Dates of marriage and schooling have to be estimated as they are based on data on current status which are collected annually, or information on ages or years of events (starting/leaving school, first marriage, start/end of marriage to specific spouses). This means that it is inappropriate to study the order of events if they occur in a short time period. The transition between states within the same domain may be studied (i.e. from divorced to remarried) but the transition between a divorce and a return to school, or a return to the family home would be difficult, because the estimated dates of the events may cause them to appear the wrong way around. Additionally, the events of divorce and returning to the family home are likely to coincide exactly. For this reason, the association between divorce and the other markers of adulthood, living with family and attending school, were examined with a cross-sectional approach. One record per participant between age 14 and 18 for women and age 16 and 22 for men was retained, only participants who were currently 'never married' or 'divorced' were kept. If a participant had more than one potential record one was kept at random. To keep the two groups independent, participants in the 'divorced' group were excluded from the 'never married' group. Two binary outcomes were assessed: living with family and attending school and the main explanatory variable was the interaction

between marital status and having at least one child. Following basic tabulations, univariate and multivariate logistic regression models (adjusted for year, parent's schooling, household occupation and the other outcome where appropriate) were run. As well as comparing the odds of each category compared to the baseline (never married, no children), the effect of children on the divorced group was assessed by testing whether the coefficients in the divorced group were significantly different from each other.

9.6 Results

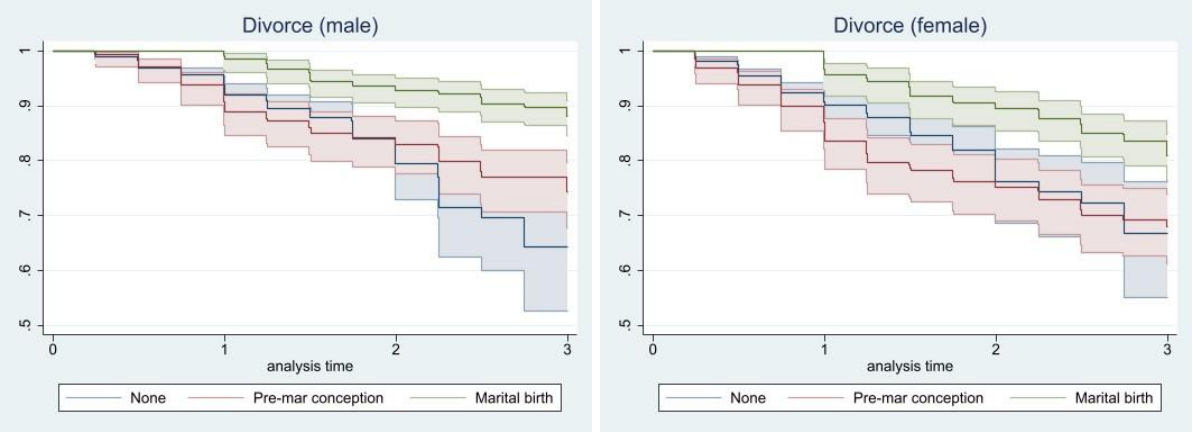
9.6.1. Rates and predictors of divorce at young age (sample 1)

The crude rates of divorce in the 3 years from first marriage were 7.8 per 100 person years (95% confidence interval [CI]=6.7-9.0) for men and 10.7 (95% CI=9.5-12.0) for women. For both sexes the crude rates were lowest if their first child was conceived once married (male rate=5.7, 95% CI=4.3-7.5; female rate=9.1, 95% CI=7.2-11.5) and highest for those who conceived a child before marriage (male rate=9.7, 95% CI=7.5-12.6; female rate=13.4, 95% CI=10.6-16.9) (table 9.1). Kaplan Meier plots of time to divorce from first marriage are shown in figure 9.1. For both men and women, those with a birth conceived within the marriage have a slower transition to divorce, though for women the confidence intervals overlap more. For both men and women, the transition appears faster for the 'pre-marital conception' group than the 'no children' group for the first 2 years, in the final year of follow-up for women the rates then appear similar but for men the 'no children' groups appears faster. For this reason, the different effects of the child categories in the 3 follow-up years were examined in the regression modelling.

Table 9.1. Crude rates by child variable and sex, of 1. divorce within 3 years of a first marriage before the age of 18 (women) or 22 (men) [sample 1]; 2. re-marriage within 3 years of a divorce before the age of 18 (women) or 22 (men) [sample 2], and 3. first marriage within 3 year of age frequency matched to sample 2 [sample 3]

	Male					Female				
	Events	Person-years	Rate per 100 py	95% CI		Events	Person-years	Rate per 100 py	95% CI	
Divorce within 3 years of first marriage (sample 1)										
None	77	8.9	8.6	6.9	10.8	83	8.0	10.3	8.3	12.8
Pre-mar conception	57	5.9	9.7	7.5	12.6	71	5.3	13.4	10.6	16.9
Marital birth	50	8.8	5.7	4.3	7.5	70	7.7	9.1	7.2	11.5
Overall	184	23.6	7.8	6.7	9.0	224	21.0	10.7	9.3	12.1
Remarriage within 3 years of a divorce (sample 2)										
No children	41	1.6	25.8	19.0	35.1	27	1.2	23.2	15.9	33.9
At least 1 child	36	2.0	18.1	13.1	25.1	62	4.1	15.1	11.8	19.4
Expecting a child	5	0.3	19.2	8.0	46.2	3	0.6	5.1	1.6	15.7
Overall	82	3.8	21.4	17.2	26.6	92	5.9	15.7	12.8	19.3
First marriage within 3 years of age frequency matched to sample 2 (sample 3)										
No children	204	23.4	8.7	7.6	10.0	350	28.2	12.4	11.2	13.8
At least 1 child	11	0.4	24.6	13.6	44.4	37	2.3	15.8	11.4	21.8
Expecting a child	21	0.2	101.2	66.0	155.1	34	0.7	46.4	33.2	64.9
Overall	236	24.1	9.8	8.6	11.1	421	31.3	13.5	12.2	14.8

Figure 9.1. Kaplan Meier plots of time to divorce (within 3 years of a first marriage before the age of 18 (women) or 22 (men) [sample 1]) by having own child, separately for males and females. NB. Scale starts from 0.5.



Results of regression analysis are shown in table 9.2. In the univariate analysis, without the interaction term, a birth conceived within the marriage was associated with lower rates of divorce for men, this effect remained in the adjusted model (adjusted hazard ratio [aHR]=0.6, 95% CI=0.4-0.9). There was very little or no evidence for an effect of duration of marriage for males or females. In the model with the interaction with duration of marriage, men had the highest rate of disruption with no children in year 3 (aHR=3.6, 95% CI=1.7-7.4). Women had a higher rate in the ‘pre-marital conception year 1’ group (aHR=1.7, 95% CI=1.1-2.6); there was also weak evidence that the rate was higher for men in this group (p=0.053). Men had a lower rate of disruption with increasing age at marriage, however there was no effect for women and there was no evidence for an effect of spouse age difference. For both men and women, the spouse being married before was associated with higher rates, however for men the evidence was weak. There was little to no evidence of effects of own or parental schooling; for men the rates were higher for those living in household with any irregular wage-earner and for women living nearer to the tarmac road was associated with lower rates of disruption. For men living with only 1 or no parents before marriage was associated with higher rates of disruption, but for women the effect was the opposite. There was no effect of living very close to own or spouse’s family during marriage for either sex.

Table 9.2. Regression model results with outcome of divorce (within 3 years of a first marriage before the age of 18 (women) or 22 (men) [sample 1]

	Male (n=1064)						Female (n=964)					
	HR	SD	Z	p	95% CI		HR	SD	Z	p	95% CI	
CRUDE MODEL WITHOUT INTERACTION												
Child												
None	Reference						Reference					
Pre-mar conception	1.0	0.2	0.3	0.800	0.7	1.5	1.4	0.2	2.0	0.048	1.0	1.9
Marital conception	0.6	0.1	-2.6	0.009	0.4	0.9	1.1	0.2	0.3	0.796	0.7	1.6
Duration of marriage (years)												
One	Reference						Reference					
Two	1.2	0.2	0.8	0.402	0.8	1.7	0.7	0.1	-1.8	0.075	0.5	1.0
Three	1.3	0.3	1.2	0.220	0.8	2.0	0.8	0.2	-1.0	0.303	0.5	1.2
CRUDE MODEL WITH INTERACTION												
Child/duration of marriage												
No children yr1	Reference						Reference					
No children yr2	1.6	0.4	1.8	0.070	1.0	2.7	1.2	0.3	0.8	0.424	0.7	2.1
No children yr3	3.1	1.1	3.1	0.002	1.5	6.3	1.1	0.6	0.2	0.820	0.4	3.1
Pre-mar con yr1	1.6	0.4	1.9	0.059	1.0	2.5	1.7	0.3	2.7	0.007	1.2	2.6
Pre-mar con yr2	1.0	0.3	-0.1	0.937	0.5	1.8	1.1	0.3	0.3	0.766	0.6	1.8
Pre-mar con yr3	1.6	0.5	1.5	0.130	0.9	2.8	1.0	0.3	0.1	0.947	0.5	1.9
Mar con yr1	1.0	0.5	0.0	0.987	0.4	2.7	2.0	0.7	1.9	0.059	1.0	4.0
Mar con yr2	0.9	0.2	-0.6	0.550	0.5	1.4	0.7	0.2	-1.5	0.127	0.4	1.1
Mar con yr3	0.7	0.2	-1.3	0.194	0.4	1.2	1.0	0.2	0.1	0.945	0.7	1.5
ADJUSTED MODEL WITHOUT INTERACTION												
Child												
None	Reference						Reference					
Pre-mar conception	1.1	0.2	0.4	0.698	0.7	1.6	1.4	0.2	1.9	0.063	1.0	2.0
Marital birth	0.6	0.1	-2.7	0.008	0.4	0.9	1.1	0.2	0.3	0.776	0.7	1.6
Duration of marriage												
One	Reference						Reference					
Two	1.3	0.2	1.2	0.236	0.9	1.9	0.7	0.1	-1.6	0.111	0.5	1.1
Three	1.5	0.3	1.7	0.095	0.9	2.3	0.8	0.2	-0.8	0.408	0.5	1.3

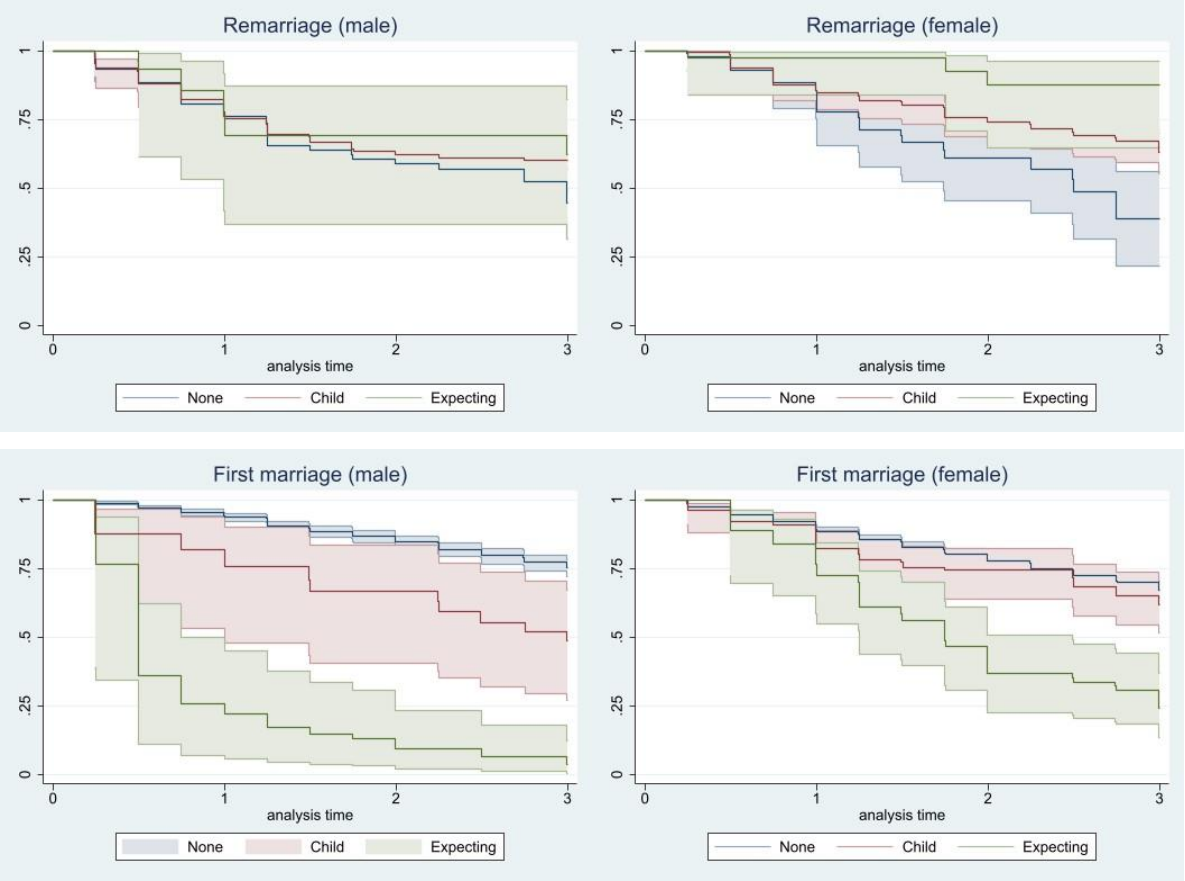
	Male (n=1064)						Female (n=964)					
	HR	SD	Z	p	95% CI		HR	SD	Z	p	95% CI	
ADJUSTED MODEL WITH INTERACTION												
Child/follow-up year												
No ch yr1	Reference						Reference					
No ch yr2	1.7	0.5	2.0	0.042	1.0	2.9	1.3	0.3	0.9	0.359	0.8	2.2
No ch yr3	3.6	1.3	3.5	0.001	1.7	7.4	1.1	0.6	0.2	0.813	0.4	3.2
Pre-mar con yr1	1.6	0.4	1.9	0.055	1.0	2.6	1.7	0.4	2.5	0.011	1.1	2.6
Pre-mar con yr2	1.1	0.3	0.3	0.793	0.6	2.0	1.1	0.3	0.4	0.695	0.6	1.9
Pre-mar con yr3	1.8	0.5	1.9	0.053	1.0	3.3	1.1	0.3	0.2	0.841	0.6	2.0
Mar con yr1	1.0	0.5	0.0	0.968	0.4	2.8	2.0	0.7	2.0	0.048	1.0	4.1
Mar con yr2	0.9	0.2	-0.4	0.685	0.6	1.5	0.7	0.2	-1.4	0.175	0.4	1.2
Mar con yr3	0.8	0.2	-0.9	0.349	0.5	1.3	1.1	0.2	0.3	0.766	0.7	1.6
Calendar year	1.0	0.0	0.8	0.417	1.0	1.1	1.0	0.0	0.8	0.447	1.0	1.1
Married age	0.8	0.0	-3.4	0.001	0.8	0.9	1.0	0.1	-0.2	0.829	0.9	1.1
Spouse age difference												
Younger	1.2	0.3	0.9	0.385	0.8	1.8						
Similar	Reference						Reference					
Older							1.0	0.2	0.2	0.867	0.8	1.4
Spouse prev married	1.8	0.6	1.8	0.071	1.0	3.5	1.7	0.3	3.1	0.002	1.2	2.3
Polygamous							1.0	0.0	0.5	0.603	1.0	1.0
Schooling												
Left prim	Reference						Reference					
Current	1.4	0.3	1.3	0.193	0.8	2.3	1.0	0.2	0.0	0.962	0.7	1.5
Comp prim	0.9	0.2	-0.6	0.548	0.6	1.3	0.8	0.2	-0.9	0.353	0.6	1.2
Some/comp sec	0.7	0.2	-1.4	0.166	0.5	1.1	1.0	0.2	0.1	0.922	0.7	1.5
Father schooling												
No secondary	Reference						Reference					
Secondary	1.0	0.2	0.1	0.889	0.7	1.4	1.0	0.2	0.2	0.876	0.8	1.4
Unknown	0.0	0.0	0.0	0.985	0.0		3.3	1.7	2.3	0.022	1.2	9.2
Mother schooling												
No secondary	Reference						Reference					
Secondary	0.9	0.2	-0.3	0.763	0.7	1.4	1.0	0.2	-0.2	0.826	0.7	1.3
Unknown	0.0	0.0	0.0	0.988	0.0		0.5	0.5	-0.8	0.453	0.1	2.8
Household occupation												
Only farming	Reference						Reference					
Any irregular wage	1.9	0.4	2.8	0.004	1.2	2.9	1.2	0.3	0.7	0.491	0.8	1.8

	Male (n=1064)						Female (n=964)					
	HR	SD	Z	p	95% CI		HR	SD	Z	p	95% CI	
Any regular wage	0.8	0.3	-0.7	0.466	0.4	1.5	0.9	0.2	-0.3	0.726	0.6	1.5
None/NK	0.6	0.3	-1.1	0.287	0.3	1.4	0.7	0.2	-1.1	0.272	0.3	1.4
Within 1km of road	1.1	0.2	0.8	0.430	0.8	1.6	0.7	0.1	-2.2	0.028	0.5	1.0
Living with single or no parent before marriage	1.5	0.2	2.5	0.012	1.1	2.1	0.7	0.1	-2.1	0.036	0.6	1.0
Current living arrangements												
Spouse	Reference						Reference					
Spouse & own fam	1.0	0.2	0.2	0.829	0.7	1.4	1.8	1.8	0.6	0.575	0.2	13.0
Spouse & inlaws	1.1	0.7	0.2	0.871	0.3	3.5	1.1	0.2	0.6	0.552	0.8	1.5

9.6.2. Effect of divorce at young age on the transition to adulthood: remarriage (sample 2 & 3)

The rate of remarriage within 3 years of divorce that occurred before the age of 22 for men was 21.4 per 100 person years (95% CI 17.2-26.6) and before the age of 18 for women was 15.7 (95% CI 12.8-19.3); in comparison the rates of first marriage in the age-group frequency matched sample were much lower for men (rate=9.8, 95% CI=8.6-11.1) but more similar for women (rate=13.5, 95% CI=12.2-14.8). The rates of first marriage were highest in the ‘expecting a child’ category, however the number of remarriages in this group was low (table 9.1). Kaplan Meier plots of the time to first marriage and to remarriage are shown in figure 9.2. For both males and females, those expecting a child have the fastest transition into first marriage, while for remarriage this group have the slowest transition (women) or are similar to the other groups (men).

Figure 9.2. Kaplan Meier plots of time to a. remarriage from divorce before age 18 (women) or 22 (men) (sample 2) and b. first marriage from age frequency matched to sample 2 (sample 3) by having own child, separately for males and females.



Results of regression analyses are shown in table 9.3. In the univariate analysis having and expecting a child were associated with high rates of first marriage for men while only expecting a child was associated with high rates for women, these effects remained in the fully adjusted model however the effect sizes were reduced (for men ‘at least one child’ aHR=2.5, 95% CI=1.3-4.6 and ‘expecting’ aHR=10.0, 95% CI 6.3-15.9; for women ‘expecting’ aHR=3.0, 95% CI=2.1-4.3). For remarriage, there was no evidence of an effect of children for men, while for women both having at least 1 child and expecting a child were associated with lower rates of remarriage (at least one child aHR=0.5, 95% CI=0.3-0.8; expecting aHR=0.2, 95% CI=0.1-0.6).

Table 9.3. Regression model results with outcomes of from a. remarriage from divorce before age 18 (women) or 22 (men) (sample 2) and b. first marriage from age frequency matched to sample 2 (sample 3), separately for males and females.

	Male (n*=215; 1024)						Female (n*=300; 1418)					
	HR	SD	Z	p	95% CI		HR	SD	Z	p	95% CI	
Remarriage												
UNIVARIATE ANALYSIS												
Child												
None	Reference						Reference					
Child	0.7	0.16	-1.5	0.121	0.4	1.1	0.7	0.15	-1.9	0.063	0.4	1.0
Expecting	0.7	0.35	-0.6	0.535	0.3	1.9	0.2	0.13	-2.5	0.012	0.1	0.7
FULLY ADJUSTED MODEL**												
Child												
None	Reference						Reference					
Child	0.8	0.20	-1.0	0.332	0.5	1.3	0.5	0.13	-2.6	0.009	0.3	0.8
Expecting	0.8	0.37	-0.5	0.594	0.3	2.0	0.2	0.11	-2.8	0.004	0.1	0.6
First marriage												
UNIVARIATE ANALYSIS												
Child												
None	Reference						Reference					
Child	2.8	0.87	3.4	0.001	1.5	5.2	1.3	0.22	1.4	0.165	0.9	1.8
Expecting	11.6	2.66	10.7	<0.001	7.4	18.2	3.7	0.67	7.3	<0.001	2.6	5.3
FULLY ADJUSTED MODEL**												
Child												
None	Reference						Reference					
Child	2.5	0.78	2.9	0.004	1.3	4.6	0.8	0.16	-0.9	0.364	0.6	1.2
Expecting	10.0	2.35	9.8	<0.001	6.3	15.9	3.0	0.56	5.8	<0.001	2.1	4.3

*First figure for 'remarriage' outcome (sample 2), second for 'first marriage' (sample 3);
**adjusted for age, year, household occupation, distance to road, schooling, parents' schooling and living with parents

9.6.3. Effect of divorce at young age on the transition to adulthood: living arrangements and schooling (sample 4)

The number and proportion of never married and divorced women aged 14-18 and men aged 16-22 according to whether they are living with family or attending school, along with results from univariate and multivariate logistic regression models and interaction tests is shown in table 9.4. A high proportion of never married men and women lived with family (85.5% and 86.3% of 'never married, no children' men and women respectively) and there

was no evidence of a difference by child status. For divorced men the percentages were lower, and divorced men with children had the lowest proportion (45.2%), and there was some evidence of an effect of child status on divorced men, however, after adjustment the evidence disappears (p-value for interaction test in adjusted model=0.276). For divorced women the proportion living with family was similar to those never married, and there was no effect of children.

A high proportion of ‘never married, no children’ men (72.9%) and women (89.4%) were attending school and the proportions were lower for those with children (51.1% for men and 29.4% for women). However, the evidence for this effect reduces for men in the adjusted model, though remains strong for women. Divorced men are much less likely to be in school compared to never married men, and while the proportion was higher for divorced men with children compared to without (17.7% vs. 11.1%) there was no evidence for this effect (p-value for interaction test in adjusted model=0.144). Divorced women are also less likely to be in school and having children reduces their chance even further (8.7% vs. 21.5%, p-value for interaction test in adjusted model=0.002).

Table 9.4. Univariate and multivariate analyses testing the interaction between marital status and having children or living with parents on the outcomes of living with parents and currently in school if never married or divorced between age 14 & 18 (female) or 16 & 22 (male) (sample 4).

	Outcome			Crude						Adjusted					
	n	y	%y	OR	SD	z	p	95% CI		OR	SD	z	p	95% CI	
Outcome: Living with family; Interaction: any child															
Male															
Model results															
Never mar, no ch	1171	6896	85.5%	Reference						Reference					
Never mar, ch	6	42	87.5%	1.2	0.52	0.4	0.693	0.5	2.8	1.7	0.82	1.1	0.254	0.7	4.4
Divorced, no ch	70	95	57.6%	0.2	0.04	-9.1	<0.001	0.2	0.3	0.2	0.04	-8.3	<0.001	0.2	0.3
Divorced, ch	68	56	45.2%	0.1	0.03	-10.7	<0.001	0.1	0.2	0.2	0.03	-8.8	<0.001	0.1	0.3
Test: Div no ch = div ch				1.6	0.39	2.1	0.037	1.0	2.6	1.3	0.34	1.1	0.276	0.8	2.2
Female															
Model results															
Never mar, no ch	1015	6375	86.3%	Reference						Reference					
Never mar, ch	8	77	90.6%	1.5	0.57	1.1	0.252	0.7	3.2	1.3	0.52	0.7	0.471	0.6	2.9
Divorced, no ch	18	144	88.9%	1.3	0.32	1.0	0.338	0.8	2.1	1.0	0.29	0.0	0.978	0.6	1.8
Divorced, ch	34	265	88.6%	1.2	0.23	1.2	0.244	0.9	1.8	1.0	0.23	-0.1	0.946	0.6	1.6
Test: Div no ch = div ch				1.0	0.32	0.1	0.933	0.6	1.9	1.0	0.33	0.1	0.943	0.5	1.9

	Outcome			Crude						Adjusted					
	n	y	%y	OR	SD	z	p	95% CI		OR	SD	z	p	95% CI	
Outcome: attending school; Interaction: any child															
Male															
Model results															
Never mar, no ch	2096	5641	72.9%	Reference						Reference					
Never mar, ch	22	23	51.1%	0.4	0.12	-3.16	0.002	0.2	0.7	0.6	0.20	-1.5	0.136	0.3	1.2
Divorced, no ch	136	17	11.1%	0.0	0.01	-11.9	<0.001	0.0	0.1	0.1	0.02	-9.6	<0.001	0.0	0.1
Divorced, ch	93	20	17.7%	0.1	0.02	-10.2	<0.001	0.0	0.1	0.1	0.03	-7.9	<0.001	0.1	0.2
Test: Div no ch = div ch				0.6	0.21	-1.5	0.128	0.3	1.2	0.6	0.21	-1.5	0.144	0.3	1.2
Female															
Model results															
Never mar, no ch	789	6633	89.4%	Reference						Reference					
Never mar, ch	60	25	29.4%	0.0	0.01	-12.5	<0.001	0.03	0.1	0.0	0.01	-12.1	<0.001	0.0	0.1
Divorced, no ch	128	35	21.5%	0.0	0.01	-17.6	<0.001	0.0	0.0	0.0	0.01	-16.7	<0.001	0.0	0.0
Divorced, ch	272	26	8.7%	0.01	0.002	-21.5	<0.001	0.01	0.02	0.01	0.003	-19.4	<0.001	0.01	0.02
Test: Div no ch = div ch				2.9	0.80	3.8	<0.001	1.7	5.0	2.5	0.70	3.1	0.002	1.4	4.3

*Living with parents/child model adjusted for calendar year, household occupation, living within 1km of main road, current schooling, parental schooling; Attending school/child model adjusted for calendar year, household occupation, living within 1km of main road, living with parent(s) & parental schooling; Attending school/living with parents model adjusted for calendar year, household occupation, living within 1km of main road & parental schooling

9.7. Discussion

9.7.1. Summary

Overall rates of divorce were higher for women compared to men. Having children affected the chance of divorce in different ways: for men a birth conceived in marriage seemed to be protective while not having children was a risk factor after the first 2 years of marriage. For women, pre-marital conception was a risk factor, mostly in the first year of marriage. There were few other independent predictors of divorce for either sex. For men, there was little evidence that divorce at a young age 'reset' the transition to adulthood: divorced men were more likely to remarry, less likely to live with family, and less likely to attend school compared to their never married contemporaries. There was also little effect of children on the outcomes of divorced men. Evidence for a 'reset' was more mixed for women: they were likely to return to family following a divorce and had similar rates of remarriage compared to never married women of the same age; divorced women, however, were not likely to attend school, and having children made their chances of remarriage and attending school much lower.

9.7.2. Rate of divorce and remarriage

Several studies have reported on the rates of dissolution of first marriage in Malawi: Bertrand-Dansereau et al found that 16.5% of their sample of rural Malawian women experienced divorce within 3 years of their first marriage (Bertrand-Dansereau and Clark, 2016); a study of DHS data which included Malawi found that about 23% of first unions dissolved in the first 4 years of marriage (Malinga John, 2022), and a study of young Malawians found that only 58% of first unions were intact by the 5th year of marriage (Grant and Soler-Hampejsek, 2014). Our rates suggest lower likelihood of dissolution (7.8 per 100 person years for men and 10.7 for women) but are not directly comparable as we include person-time for people who leave the area and thus may be an underestimate if we have missed marriages that occurred elsewhere; also the other studies included older people and covered time periods not included in our study.

Remarriage has been found to be common and often rapid, with an average time between union dissolution and remarriage in Malawi of 1.9 years (this included all unions at all ages, not just first unions) (Malinga John, 2022), so it is not surprising that we found high rates of remarriage following divorce. Our finding that men had higher rates of remarriage than women also confirms previous findings that in general remarriage has been found to be more common for men than women (de Walque and Kline, 2012). It has also been found

chance of remarriage for women decreases after age 19 while remaining stable for men at all ages (Reniers, 2008), which further explains the sex differences. Other factors previously found to inhibit remarriage are widowhood (Reniers, 2008), HIV status and living children (Anglewicz and Reniers, 2014): very few of the young people in our sample were widowed and we do not have HIV status data, but we demonstrated that having children was associated with lower rate of remarriage for women.

9.7.3. Predictors of divorce

The association between pre-marital conception and a higher chance of divorce for women has been found previously (Odimegwu et al., 2017; Smith-Greenaway et al., 2021) and maybe due to the pregnancy 'forcing' the marriage in young people who perhaps have not known each other very long or are not otherwise ready for marriage, as was found in a qualitative study of young people in Zambia (Mweeba and Mann, 2020). It has also been found that pre-marital conception was associated with higher likelihood of experiencing a violent marriage, regardless of whether it ended in divorce or not (Smith-Greenaway et al., 2021).

Lack of children in a marriage has previously been shown to be associated with higher chance of divorce for women (Bertrand-Dansereau and Clark, 2016; Reniers, 2003; Takyi, 2001; Tilson and Larsen, 2000). We only found an association between divorce and lack of children for men, not women. This is somewhat surprising as many of the same couples are likely to be in each group, however some of the women maybe marrying older men who may be in more of a position to marry another woman, if infertility is suspected, rather than initiating divorce (Hemmings, 2007). The young men in our analysis may not yet be financially stable enough to attract a second wife.

The evidence on the associations between socio-economic position and divorce were not totally clear in our analysis. We did not find an effect of education, however the findings in previous literature have not been consistent with increased education in women associated with lower chance of divorce in Southern Malawi (Grant and Pike, 2019; Grant and Soler-Hampejsek, 2014) and also higher chance of divorce in Ghana (Takyi and Broughton, 2006). We found that for men living in a household in the 'any irregular wage' category was associated with higher rates of divorce. Having to rely on irregular work could be an indicator of lower household wealth: other research has shown that low household wealth in the marital household was associated with higher chance of divorce in Uganda (Porter et al., 2004), and qualitative accounts in Zambia described how marriages may be disrupted in

young people if the husband did not have the resources to take care of his wife (Mweeba and Mann, 2020). We found an association between older age at marriage and lower chance of divorce for men, but no evidence of this in women, which is somewhat surprising as this has been found to be a risk factor for divorce for women in other studies (Clark and Brauner-Otto, 2015; Reniers, 2003; Tilson and Larsen, 2000). We also did not find effect of spousal age difference, while it has been found previously that women who married men older than them were less likely to experience divorce (Grant and Pike, 2019; Porter et al., 2004; Reniers, 2003). We were not able to examine different age differences do to relative small sample size: the associations may have been different for larger age gaps. All of our index young people were married for the first time in this analysis, but the increased chance of divorce if their spouse had previously been married confirms that prior divorce is a predictor of future divorce (Porter et al., 2004; Takyi, 2001). It has been should previously that polygyny is associated with higher chance of divorce for Malawian women (Reniers, 2003), however we did not find any evidence of this in our data.

Living with only 1 or no parents prior to marriage was associated with lower rate of disruption for women, which might be expected if the decision to leave a marriage depends on the home that a young woman has to go back to, and that certain household types might be more receptive to accepting the woman back. Mweeba and colleagues in Zambia found that young women who left a marriage always had relatives to go back to (Mweeba and Mann, 2020), however a Malawian study suggested that it was the payment of the lobola (bride price) from the husband to the wife's family that influenced whether or not the family would accept her back (Bertrand-Dansereau and Clark, 2016). We did not have information on the payment of lobola in our data. This same study however found that the couples that separated tended to be those who lived independently from their families following marriage, (Bertrand-Dansereau and Clark, 2016) however we found no association between divorce and living near own or spouse's family. For men, the association with pre-marital living arrangements was opposite: they were more likely to divorce if they had not been living with both parents. It has previously been found that parent's death and divorce was associated with their child's later chance of divorce, though this study was only in women (Grant and Pike, 2019).

9.7.4. Effect of divorce on transition to adulthood

Studies on the effect of divorce on young people's transition to adulthood in Africa are rare and not directly comparable to the present analysis. A study in Malawi found 20% of young women following a divorce were living independently and that they were at a disadvantage in terms of material well-being, however they did not compare to a never married group.

Although the authors suggested that those returned to their parents may be able to ‘reset’ the transition to adulthood, they did not look at attending school (Grant and Pike, 2019). Another Malawian qualitative study on implementation of marriage age laws and marriage withdrawal reported that while often there were negative connotations of being withdrawn from marriage some people said that girls could potentially go back to school, but there was no evidence presented that people did actually return to school (Melnikas et al., 2021).

We have shown that divorced men and women are different from never married men and women of similar ages, according to 3 common measures of the transition to adulthood: moving away from home, leaving school and getting married. Individual’s transition to adulthood may be complex and vary by order and timing of events, for simplicity’s sake however consider 2 extremes: an early marriage route and an extended education and later marriage route. The never married group in our analysis have the potential to follow either route (plus other routes), while the divorced group have started on the early marriage route. If the levels of schooling and marriage in the divorced group were more similar to the never married group it could mean that some had switched the route they were taking following the divorce to get more education before remarrying. Removing the effect of children initially by only considering divorced men and women without children: their higher rates of marriage compared to their never married counterparts and relatively lower likelihood of school attendance implies that most are not deviating from the pathway (early marriage) that they began at an earlier age. The lack of the association between expecting a child and remarriage rates compared to the very clear association with marriage rates, however, may show that the societal/family pressure to marry in this situation has disappeared, or possibly that divorced people may be more careful to avoid pregnancy. Divorced women with children have lower rates of remarriage and of school attendance which may mean that both continuing on their existing adulthood trajectory or switching to an extended education one is more difficult for them. Without information on the intentions and desires of the participants the interpretation can only be speculative: the women with children may neither want to remarry, live independently or finish school, however in Zambia, divorced men and women expressed desires and hopes of finishing education and many young women resented the loss of independence after returning home following divorce (Mweeba and Mann, 2020). Finally, divorce may not ‘reset’ the transition to adulthood, because those who marry early may have different characteristics than those who continue in education. Individuals from less wealthy households, who may not be able to capitalise on any opportunities which may arise from extended education are likely to be those who marry early.

9.7.5. Limitations

Our secondary longitudinal data allow for detailed analyses of the early marital experiences of young people in rural Malawi, it is valuable firstly as data are available for both women and men, as quantitative data on divorce in men are limited; and secondly as studies often use retrospective data gathered from older people: a study using data from Malawi Longitudinal Study of Families and Health found that marriage likely to be missed or incorrect dates reported in retrospective data collection, this was more likely for short marriages or those a long time ago (Chae, 2016);

However, the data were not specifically collected for this purpose, which has resulted in lack of data which might have informed this analysis, and having to make some assumptions which may have weakened the associations observed. The HDSS maintains surveillances over households and participants in a particular area: no data are available on people when they are outside of the area. As marriage and divorce are common reasons for moving in and out of the area, especially for women, we have attempted to include these outcomes. However, due to the way the data are collected the reason for moving may refer to another person (i.e. an adolescent may move due to their parents' divorce): to reduce this effect only independent moves were classed as marriage or divorce outcomes, however there is still the possibility of misclassification.

Data that would have been useful to include but was not available include HIV status: this has been found to be important for both divorce and re-marriage (Anglewicz and Reniers, 2014; Porter et al., 2004), while the Karonga HDSS has carried out HIV sero-surveys (finding an adult [18+] prevalence of 7.5% in 2007-8 (Molesworth et al., 2010)), these tended to include only adults, and there was not enough data on all of our young people to assign HIV status; type of marriage: traditionally in Malawi marriages are negotiated by 'ankhoswe' who are senior relatives on either side, involve a traditional ceremony and the payment of 'lobola' or bride price from the husband to the wife's family. It has been previously found that elopement (where some or all of these traditional aspects are avoided) is associated with higher chance of divorce (Bertrand-Dansereau and Clark, 2016; Grant and Pike, 2019), however this information is not available in our dataset. The annual nature of household and individual surveys means that it is necessary to assume that a report of, for example, occupation, is true for a certain time after it is made. This also meant that we assumed that the absence of the spouse in the household meant that they were separated, even if they were still assigned the status of married, as marital status is only gathered annually. Absence of spouse may mean that they are working elsewhere or living away for a reason other than divorce: this has been

shown to be relatively common in other parts of rural Malawi (Reniers, 2003), however qualitative studies have shown that temporary separations regardless of the initial reason can lead to divorce in young people (Bertrand-Dansereau and Clark, 2016). Our main explanatory variables of child status were deliberately kept as simple as possible due to the relatively small numbers in some of the groups: it has been suggested number and sex of children can affect divorce in Africa (Odimegwu et al., 2017), however as our timescales were relatively short it seems unlikely that child sex would have a big impact.

The focus of the analysis was on divorce at a young age: in choosing the age ranges for the inclusion criteria it was necessary to balance trying to look at the youngest age groups with having enough data to have statistical power in the analyses. It may have been preferable to use younger age cut-offs for the divorce analysis, i.e. include only women married by age 16 and men by age 20, to truly focus on those marrying very young. This group was much smaller but the estimates from the models were similar to those using the expanded sample but the evidence was weaker, so it was decided to use the same age restrictions for all the analyses. It was necessary to use different age criteria for men and women, as women tend to marry earlier than men in the area. This makes it harder to compare the results between the sexes.

9.7.6. Conclusion

We find evidence that both men and women are affected by having children with regard to first marriage and divorce, and that a divorce at a young age did not appear to change the transition to adulthood trajectory for men, or women without children. However, divorced women with children may be at more of a disadvantage for both remarrying or returning to school.

10. Discussion & conclusions

In this discussion I first highlight key results and conclusions from the thesis, specifically relating back to the overall objectives stated. Secondly, I pull together aspects pertaining to adolescence and the transition to adulthood, drawing conclusions and identifying areas for further research. Thirdly, I use the same framework I used in the literature review in chapter 4, to discuss issues related to data manipulation and analysis, including strengths and limitations, and important areas to consider in future work. Fourthly, I consider how similar analyses could be carried out with data from other HDSSs. Finally, I summarise identified areas for further work, and my overall conclusions.

10.1. Summary of findings in relation to thesis objectives

My overall objective was to demonstrate the use of complex data manipulation and reduction techniques on existing data from the Karonga HDSS to usefully answer questions related to health and demography. I will use this section to summarise how this has been achieved.

Chapter four presented the results of a literature review of analyses which leveraged key linkage aspects of HDSS datasets, this was used as a vehicle to explain the issues which arise from using data and way in which to deal with them. I identified an array of different types of analyses which used lots of different types of data manipulation, data structures and statistical techniques, demonstrating the utility and value in for complex secondary analyses using HDSS data (objective 1a). The papers used some different techniques to account for HDSS-specific issues such as migration, which can introduce bias in some analyses either due to it being related to the outcome or through data inconsistencies and missing data which may affect participants differently according to how long they are present in the study. I made some suggestions to current and potential users of HDSS data on how best to conduct and present analyses to account for these issues. This paper will increase knowledge and visibility of HDSS datasets and the scope of potential secondary analyses (objective 1c).

In chapter 5, I present the dataset which I created to be the base dataset for all my analyses. I describe the data manipulations, the limitations and possible uses of the data. By describing the above in quite some detail, this paper will also increase knowledge and visibility of HDSS datasets and the scope of potential secondary analyses (objective 1c),

either by encouraging collaborative usage of the Karonga data, or encouraging other HDSSs to consider whether similar datasets could be created.

In chapter 6, I assessed the value and potential disadvantages using latent class analysis (LCA) with HDSS data (objective 1b). The LCA produced household composition variables which demonstrated the variety of household types where adolescents were living (objective 2a). The data-driven approach generated more categories than the 'traditional' variable, which divided households into 'nuclear', single parent, 'blended' and 'extended', suggesting the need for greater flexibility when categorising households i.e. to distinguish between 'extended' family types, i.e. maternal and paternal. Additionally, the approach suggested that 'nuclear'-type households, which I termed 'parents and siblings', should have a more flexible definition, including those with some other relatives (as well as both parents). Using the expanded household shed further light on the way that living arrangements can be described: being able to distinguish between single parent households with and without (potentially supportive) grandparents, or other family, nearby may provide valuable insights for analyses. The analyses showed that while 'parents and siblings'-type households accounted for a large proportion of adolescents' households, a good number were living in other types. Using the expanded household definition reduced the number living in 'parents and siblings'-type households even further, with a proportion living near older brothers, paternal family or in a polygamous set-up. I found LCA useful for generating household composition groupings. However, for repeatable analyses using the longitudinal HDSS data, I found it most useful to use as a guide for developing manually created categories (see below sections for a full review of using LCA and related techniques with HDSS data).

The sequence analysis chapter (chapter 7) describes the transition to adulthood in young Malawian women (objective 2b), demonstrating the use of HDSS data for secondary analyses not initially planned (objective 1a). I found that, for many young women, the transition to adulthood follows quite a traditional route: leaving school and home to marry, then rapidly building a family. There was, however, a sizable group who remained unmarried and attending school during the analysis period, and this group seemed to become more common over time. I considered the value and potential disadvantages of using sequence analysis with HDSS data (objective 1b), in particular how the results are affected by migration. Attempting to include data from individuals without complete sequences (because they had left the area) did not change the initial conclusions dramatically. However, it did highlight the relatively large group of women who left the area before being married, whom we cannot draw conclusions about. It also highlighted a group of women who experience divorce early in their marriages. Using sequence analysis on this dataset with a relatively

high proportion of missing data was somewhat challenging, but still produced useful results, especially as there are few other sources to examine the transition to adulthood in low and middle income countries. I found sequence analysis to be very helpful for exploring and visualising the data to help guide analytical question development.

The migration analysis in chapter 8 followed from the LCA in chapter 5, assessing how often young people changed household and whether there was an influence from the presence of family outside of the household (objective 2a). I showed that moving is common for young people, especially over adolescence when both males and females move a lot independently; however, adolescents also move accompanied by the families so not all moves are related to the transition to adulthood. There were clear differences by sex, with differences in migration rates from as young as 4 years old: the peak of movements over adolescence occurred several years earlier for girls than boys, and girls and young women tended to move further. Presence of family locally was associated with lower chance of moving: the maternal family seemed important in particular. There did not seem to be evidence of a high level of movement of young people to urban areas: this helps to be more confident in the results of the transition to adulthood sequence analysis, implying that the group that leave the area are probably not pursuing a very different path to adulthood through work and education in the city, but are moving to other rural areas. This analysis demonstrates the value of using HDSS data for secondary analyses (objective 1a) and the utility in complex data manipulations (objective 1b), in particular, the linkages within households and families allowed me to categorise moves as ‘independent’ and ‘accompanied’ which provided very useful nuance in terms of using the data to understand the transition to adulthood.

The analysis of divorce and remarriage at a young age in chapter 9 provided further insight into the group of participants who may not have a smooth transition to adulthood (objective 2c). Divorce was found to be quite common, and influenced both by having children and the timing of conception (whether pre-marital or not); although for men, a lack of children by the 3rd year of marriage was associated with higher likelihood of divorce. Divorced men were likely to remarry quickly and not return to school following divorce, and their outcomes seemed largely unaffected by whether they had children. Women were slower to remarry, especially if they had children, and were also not likely to return to school: this was especially the case if they had children. This analysis also demonstrates the value of using HDSS data for secondary analyses (objective 1a) and the utility in complex data manipulations (objective 1b) as the longitudinal linkages, and linkages between spouses allowed for identification of transitions between various marital states.

10.2. Adolescence and the transition to adulthood

Adolescence is a key time in an individual's life, their experiences, including their living arrangements, during this time and the transition to adulthood can have effects on their well-being and health in the short-term, but also can have far reaching effects in their lives, and possibly the lives of their children (Delprato et al., 2017). The analyses in this thesis have shown the fluidity and complexity of adolescence in Malawi. Adolescents live in a range of different household/family structures, are relatively mobile, and experience a number of different transitions including in and out of marriage. On the other hand, adolescents seem to be experiencing these changes within the relatively stable context of a family network, which often includes family members outside the nuclear family. The longitudinally linked HDSS data allowed me to examine the lives of the adolescents in quite some detail, and use of data driven techniques goes someway to avoid interpreting the data through a Western bias, and gives the potential of detecting changes in behaviours and trends without specifically looking for them.

The majority of adolescents were living in households that included members of their nuclear family (though not necessarily both parents), but a good section were living with either maternal or paternal family. There have not been in-depth analyses of adolescent living arrangements in Malawi before, though a study of female adolescents from DHS data found that about 35% were not living with either parent (Shoko et al., 2018). In my data about 19% were living with neither parent, however the Shoko study only included girls aged 15-17 while my analysis included both sexes aged 12-18. The sequence analysis suggested that there was a group of adolescent girls not living with their parents but attending school, rather than marrying. However, one of the example analyses from the LCA paper suggested that schooling outcomes for adolescent girls were worse if they were living in 'maternal' households, compared to 'parents and siblings'. In this area, children may go to live with other relatives to be able to attend school. However they may also be living without their parents because of orphanhood or helping the other relatives, which might make it less likely for them to attend school: a Malawi analysis found that among children not living with at least one parent, only those who were double orphans had poorer school outcomes compared to children living with at least one parent (Hampshire et al., 2015). Further analyses of the Karonga HDSS data would be required to fully understand the relationship between family, living arrangements and schooling.

Time constraints prevented a thorough analysis of household composition types by sex, though in some initial exploratory analyses I did find some evidence that girls were more likely to live with the ‘female’ side of the family, with sisters or maternal relatives, and boys with ‘male’ side, with brother or paternal family. The migration analysis did allude to this somewhat, as females were more likely to move to maternal households, and were more likely to be accompanied by their mother, compared to males. Gender differences in living arrangements have been studied in Africa, however the focus is usually in older people (Kimuna, 2005; Schatz et al., 2018). There does not appear to be much in the literature regarding adolescents, though one South African study of children up to age 18 did mention that no substantial gender differences were apparent (Madhavan et al., 2017b). A full analysis of the Karonga HDSS data (of all ages), perhaps with some comparisons with other countries, could give some unique insights into important cultural issues. Other aspects of the gender divide have been shown through my analyses: the LCA household analysis looking at the ‘expanded’ household definition showed that when both parents were present then, if they were living in close proximity with family, it is likely to be paternal or a brother’s family. Adolescents were only likely to live very close with maternal family if their father was not present in the household. The distinction between maternal and paternal family is not always made in family/household research in Africa, however when it is used interesting nuance is revealed. For example, a South African study of changes in living arrangement for children found households including maternal kin to be more common and increasing by birth cohort at a higher rate than households including paternal kin (Madhavan and Brooks, 2015). Another study on young adults in a different area of South Africa suggested that available kinship-ties had shrunk over time, particularly on the paternal side (Harper and Seekings, 2010). Further analysis of the Karonga HDSS dataset, especially longitudinally, may reveal further nuances about these relationship types.

I showed some evidence for change in the transition to adulthood in this area of rural Malawi, towards increasing likelihood of later marriage, however the group of women who married early was still sizable in the later birth cohort. Evidence from the Malawi DHS shows that while overall mean age at first marriage has remained quite stable, in the north (where the Karonga HDSS is located) it has increased, corroborating my findings (Baruwa et al., 2019). I did some simple analyses testing the associations between sequence group and socio-economic status (SES) which I did not include in the thesis chapter, but presented at the International Population Conference in 2021: finding that the ‘no/late marriage’ group were more likely to be in the higher SES group, the earlier marriage group more likely to be in the middle or high group, the ‘divorce’ group more likely to be in the lower SES group, and the group who ‘migrate for marriage’ to be in the lower SES group (McLean et al., 2021b).

The same analysis of Malawian DHS data found that mean at age marriage was highest for those with more schooling and in the higher wealth quantile (Baruwa et al., 2019). These associations need to be explored in greater detail, as low socio-economic status has been linked with both fast transition to adulthood, i.e. school drop-out, marriage and child-bearing in young women who may not be physically or psychologically ready (Palamuleni, 2011); and also slow transition to adulthood, i.e. economic conditions leaving African youth unable to begin their independent lives (Hownana, 2012).

My divorce paper provided evidence that once on a particular type of transition to adulthood pathway, young people do not tend to change: i.e. the breakdown of a marriage is not likely to put someone towards the education path, even for young men, who tend to be unencumbered by child-care responsibilities in this area. I only used data up to 2017, before there was a crack-down on child marriage in Malawi, resulting in girls in some areas being removed from marriages, and returned home with the aim of them going back to school (Melnikas et al., 2021). I was only able to use data up to 2017 because there were gaps in Karonga HDSS data collection from 2018 to 2021 due firstly to funding issues, and then to the COVID-19 pandemic. Now that data collection is continuing, when possible, it will be useful to look at whether the rates of divorced young women returning to school has increased recently. While ending 'child' marriage is ostensibly a good thing, in rural communities with limited educational and employment prospects it is important to consider the relatively arbitrary age cut-off of 18 with a little more nuance. A paper drawing on qualitative and quantitative data from Tanzania argues that focussing on this arbitrary age, with the insinuation that all marriages in younger girls are forced and damaging, neglects to take into consideration the views and experiences of the women involved. While actual forced marriages are of course unacceptable (and may occur after the age of 18 as well), for a young woman with limited access to decent schooling and few job opportunities, entering into a marriage at a relatively young age may be her own choice, and the best option for her. Waiting until the age of 18 may not mean gaining any education or social advantages, or could actual put her at a disadvantage if she faces stigma from the community (Schaffnit et al., 2019).

The multi-channel sequence analysis showed that women only really left home for marriage, so I designated this the 'key marker of adulthood' focussing on it for the remainder of the sequence analysis chapter and investigating it in more detail in the divorce paper. Further evidence for this was from the migration paper, where the majority of independent moves for adolescents of both sexes was from a family home to a spousal one (or vice versa for some female adolescents). Of course, as in all of my analyses, I have mostly been reliant on

conclusions I can draw from the data. I am fortunate to be familiar with the HDSS area, spending plenty of time there over the course of my PhD (and before) and having easy access to insights from Malawian staff living and working in the area. However, it would be useful to carry out qualitative studies with the adolescents in the area to see whether this assumption is fair, and whether the social markers of adulthood are changing.

10.3. HDSS data issues

In my literature review in chapter 4, I identified 6 aspects which I used to review the studies: data manipulation techniques, dataset structures, statistical methods, repeated measures, migrations, and missing data. I will now use the same aspects (albeit with some headings combined where appropriate) to discuss some of the issues highlighted in my thesis, including strengths and limitations of the Karonga HDSS dataset, and aspects which relate to the wider HDSS community.

10.3.1. Data manipulation techniques & dataset structures

From the literature review, I identified 10 forms of data manipulation divided into 2 main categories, individual linkages over time and links between individuals. Through the course of my analyses, I used most of the types of data manipulation and linkage identified. In table 10.1 I have listed the main data manipulation techniques I have used, with the type of data manipulation/linkage used: I have included the manipulations from the core dataset preparation described in the methodology chapter with each analysis as appropriate. There were two types of data manipulations that I identified in the literature review but did not use, summary measures not linked to a date, and matched analyses, so these do not appear in the table. Most of my data manipulations made use of individual linkages between parents and children, spouses and household members to create variables which were used as exposure variables, often time-varying, for analyses of outcomes. I have also used linkages across time between individual's data points to generate outcome events, lagged exposures and sequences of events. My unit of analysis was always the individual, however I made some household summary variables, which again were used as exposures for individual analyses. I did not take advantage of the possible linkages to conduct matched analyses, nor examine household-level events, though both would be possible using the data. These manipulations led to a variety of dataset structures, in the LCA household analysis I reduced the data to a cross-sectional snapshot; in the sequence analysis I first expanded the continuous episodic HDSS data to one record per quarter, and then reduced it back to one record per person with multiple variables to create each sequence; in the migration analysis I

used a quarterly panel dataset derived from the episodic HDSS data; and in the divorce analysis I transformed the quarterly dataset back into an episodic format, for a more traditional time to event analysis.

Table 10.1 Listing of data manipulation processes for each analysis

	Linkages across time				Linkages between individuals			
	event date	time-varying	sequence	Lagged	Single link	HH summary	Other summary	HH event date
LCA Household composition								
Individuals linked via household, parent and spouse IDs to create household composition variables					x	x		
Link individuals to child records to identify timing of child-bearing, and linkage over time to create outcome of 'birth next year'				x	x			
Sequence Analysis								
Individual linkage over time to create sequences			x					
Link individuals to child records to identify timing of child-bearing					x			
Individuals linked via spouse ID and household ID to confirm if living with spouse for marital status					x			
Individuals linked via household, parent and spouse IDs to create household head variables					x	x		
Migration analysis								
Individual linkages over time to identify migration event	x							
Individuals linked via household, parent and spouse IDs to create household composition variables		x			x	x		
Individuals linked via parent IDs to create variables for family living within 250m		x			x			
Individuals linked via parental & HH IDs to identify if moves are independent or accompanied					x			
Linkage via HH ID to generate household age composition variable		x				x		
Linkage via parent ID to generate orphanhood and parental education variables		x			x			
Linkage of households within certain geographical area to generate population density variable		x					x	
Linkage via HH ID to generate household head employment score		x				x		

	Linkages across time				Linkages between individuals			
	event date	time-varying	sequence	Lagged	Single link	HH summary	Other summary	HH event date
Divorce analysis								
Link individuals to child records to identify timing of child-bearing		x			x			
Individual linkages over time to identify dates of marriage, disruption and remarriage	x							
Linkages between spouses to identify spouse previous marriage, spouse age difference and polygyny					x			
Linkage via parent ID to generate parental education					x			
Linkage between household ID to generate household occupation variables		x				x		
Individuals linked via household, parent and spouse IDs to create living arrangements variables		x		x	x			

On top of the data manipulations used to create the variables and format the data into the structures needed for each analysis, I also carried out data cleaning to deal with missing data and inconsistencies (see separate section below). I use Stata for all my data manipulations, and I always attempt to document my do-files. However, understanding another person's code can be a challenge, even for someone experienced with the same software, meaning that it is difficult for others to reproduce analyses. There has been progress in recent years towards greater transparency in science, with more journals and funders requiring that datasets and programming files be made available, ideally as open access. For the submission of the migration paper to Wellcome Open Research I made my data manipulation and analysis coding files available (https://zenodo.org/record/7797357#.ZCvY6&%2395;bMJD8). However, just sharing code does not fully address the problem as, firstly, knowledge of Stata is needed to fully understand the manipulations. Secondly, many of the decisions leading to specific commands or sets of commands are contingent on the data structure at the time, on earlier analytical decisions and on data collection structures, it is therefore hard to understand the implications of specific sections in isolation from all the other steps. There are documentation standards for static datasets (https://ddialliance.org/) but there is currently not an equivalent for the process used to produce the datasets: this limitation is recognised in the documentation community.

For some of our other datasets at MEIRU we have attempted to document the data transformations used. Table 10.2 shows two examples of derived variables from data from a long-running TB surveillance study. The variables involved and their original tables are listed, plus a description in words followed by the actual code. The first requires a two-step data processing, while the second is relatively simple; however, it uses multiple variables from multiple tables: the individual tables and how they are initially combined is described separately. This requires extensive input from the person writing the commands to recode in words what was done, and keep the comments updated with any changes. The comments then need to be transferred manually to the documentation program. This process is not particularly sustainable. Additionally, it still requires understanding of the Stata language. There have been recent developments in automating this process. The Continuous Capture of Metadata for Statistical Data Project (C2Metadata) has developed an automated system that 'reads' code used to manipulate and transform data (currently Stata, R, SPSS, SAS and Python) and 'translates' it into a human-readable form that does not require any knowledge of the source program. They have called this form the Structured Data Transformation Language (SDTL), and the system automatically adds the information on derivation to the metadata for the variable (Alter et al., 2020). Several aspects of documenting data transformation of HDSS data, including using of SDTL, were examined in detail by my colleague Chifundo Kanjala in his PhD thesis (Kanjala, 2020). Further developments in automating these processes should help HDSS data producers and users. Firstly, it would help HDSS data producers to be able to combine their data as it would be easier to assess whether specific variables relate to the same or similar concepts; secondly it would help potential data users (whether experienced with other HDSSs or not) use an HDSS dataset appropriately. Any 'human-readable' documentation for a complex data process whether manually or automatically produced, will still be relatively complicated and require skill and experience to use. An useful example of this is the wealth index documentation for DHS: documentation detailing how this one variable is created takes the form of a 7 pages of text, followed by a further 70 pages of syntax examples from SPSS (https://dhsprogram.com/programming/wealth%20index/Steps_to_constructing_the_new_DHS_Wealth_Index.pdf). Having a standard format for data process documentation would make understanding interpretation much easier.

Table 10.2 Example manually transcribed documentation of data transformations from another study

VARIABLE	spstdy: In Spouse study	schlever: Ever been to school
PROCESSING	variable: idcase [ident of TB patient] (table[s]: tb_tbs1) variable: intdate [interview date] (table[s]: tb_tbs1) Variable is "yes" if record exists for this episode on spouse table <the spouse data are put together separately first:> use ${statafiles}\tb_tbs1, clear keep idcase intdate rename (idcase intdate) (ident spdate) bys ident (spdate): gen n=_n sum n local spmax=r(max) reshape wide spdate, i(ident) j(n) save ${tempfiles}\tbspousewide, replace <and then merged into the main file:> merge m:1 ident using ${tempfiles}\tbspousewide, keep(master match) nogen gen spstdy=0 forvalues x=1/`spmax' { bys ident (intdate): replace spstdy=1 if spdate`x'[_n]>=starteps[_n] & (spdate`x'[_n]<starteps[_n+1] | _n==_N) & case[_n]==1 & spdate`x'~=. }	variable: schlever [ever been to school] (table[s]: tb_tbx, tb_tbo, tb_tboto2007) variable: attdschl [currently attending school] (table[s]: tbold_tbpart1/2, tbold_control) variable: prevschl [previously attended school] (table[s]: tbold_tbpart1/2, tbold_control) Variable is updated to include data from older versions of schooling variables replace schlever=1 if attdschl==1 replace schlever=2 if prevschl==1 replace schlever=0 if attdschl==0 & prevschl==0

I have not made any of my datasets for my papers available open access, partly because the very detailed information about family and dates of events could make individuals identifiable. Work has been done to develop methods to anonymise HDSS data which reduces the chances of individuals being identifiable yet maintaining the utility of the dataset (Templ et al., 2022) and to ‘geomask’ datasets with GPS data so that privacy is maintained but useful analyses can still be carried out (Hunter et al., 2021). These methods, however, are not in standard usage, nor do they cover the extent of variables in my datasets. The other reasons for not making my datasets available open access are, for the reasons described in this thesis. Firstly, care and guidance are needed to appropriately analyse

them. Secondly, to maintain some control over who can analyse them to ensure that paper-writing opportunities and credit are given to the people involved in capturing the data. I will publish my methodology chapter as a ‘data note’ with the aim of providing potential collaborators with information about the specific dataset, and also increasing visibility of MEIRU data in general.

10.3.2. Statistical methods

In my thesis I used a range of analytical techniques: latent class analysis in the household paper; sequence analysis and cluster analysis in the sequence analysis chapter; multi-nomial multi-level modelling in the migration paper; descriptive Sankey diagrams in the household paper and the migration paper; event history analysis or survival analysis in the marital disruption paper; and logistic regression in the household paper and the marital disruption paper. The regression modelling techniques I have used are commonly used for HDSS data so require little discussion, however I will discuss some aspects of the other techniques below.

One of the advantages of longitudinal continuous HDSS data that is available over a long period is the high number of possible ways to look at the data. In most of my analyses I reduced the continuous data to quarterly snapshots; however any period of time could be chosen, assuming that the exposure or outcome data are collected with enough frequency. The good quality linkage to parent and spouse IDs and GPS data allowed the identification of relatives of any kind, living at any distance at any point in time, or sequences of time-points, expanding the potential data available to almost overwhelming levels. Various data reduction techniques exist to help make large datasets more manageable, and I used two in this thesis: latent class analysis (LCA) and cluster analysis (CA) (following sequence analysis).

LCA and CA are similar in the sense that they use algorithms to partition records into groups based on the information inputted: in my case for LCA this was the series of binary or categorical variables indicating presence of relatives in the immediate or expanded household, and for CA this was the dissimilarity matrix generated by the sequence analysis. In both techniques statistical and theoretical criteria are used to choose the number of groups, however they are different in that with CA observations are assigned to one group, but with LCA each observation has a probability of membership of all groups. The latter can sometimes lead to difficulties in assigning observations to groups if they have similar

probability of membership of more than one group, however, equally this does serve to demonstrate that the model has not worked well (Weller et al., 2020). Both techniques also require user-input into naming/describing groups, which may artificially cover some of the complexity within the group (Weller et al., 2020). I used CA in the sequence analysis chapter as it tends to be the standard way of partitioning the distance matrices following sequence analysis, although, as mentioned in the sequence analysis chapter there is still debate over whether other techniques would be better (Liao et al., 2022). I could also have used CA for the analysis of household compositions, however I opted for LCA as my assumption that there was a 'latent' household composition variable to detect seemed to fit the technique better.

Learning LCA was also useful for me as it belongs to a group of techniques that also include longitudinal methods. Latent transition analysis (LTA) calculates the probabilities of transition from 1 class to another over 2 or 3 time points; latent class growth analysis (LCGA) uses just one indicator variable which has repeated measures over time and identifies groups of growth/change trajectories; and repeated measures LCA uses multiple indicators but over time. It does not require change over time however, which for certain topics can make it more useful than LCGA (Killian et al., 2019). In this thesis I conducted the LCA on data from multiple years as repeated cross-sectional analyses, rather than longitudinally. This was because I was primarily interested in initially describing the households, rather than examining changes over time. Some of the longitudinal methods may be useful for use with HDSS data, however the issues with incomplete data due to migrations, described in the sequence analysis chapter, would also need to be taken into account.

LCA and CA have been used with HDSS data before mostly also on cross-sectional data. For example a study from Nairobi HDSS used LCA to categorised patterns of sexual behaviour from a cross-sectional survey of HDSS residents (Maina et al., 2020); and a study from Cuatro Santos HDSS in Nicaragua used cluster analysis to generate groups of multi-factorial poverty (Källestål et al., 2020). Two studies did use LCA on longitudinal data. The first looked at sexual behaviour over time, but this was based on retrospective data collected in one survey (Angotti et al., 2018). The second used 3 rounds of prospectively collected data on sexual partners and conducted LCA on the information on the partners to generate a partner type variable. However, the 3 rounds of partner data were pooled and one LCA was run on the combined data: there was no way to identify if the same partners were reported by more than one participant, or in more than one round, and no attempt to allow for attrition (i.e. participants who provided data for more rounds might contribute more partners who might be different to partners of participants who dropped out of the survey (Nguyen et al.,

2019). For my work, I did consider doing the LCA on one large sample including all the households from all the years, rather than separate LCAs per year. The reason I did not was that stable household types would be over-represented in the sample which might then produce confusing results. For my work, repeating the LCA in the separate years demonstrated the stability of the technique on slightly different datasets. Additional, carrying out the LCA on each year would allow for emergent household types to be detected, which may be drowned out in a pooled analysis.

Other data reduction techniques that could be useful for HDSS data are those grouped within the factor analysis family, for example principal component analysis (PCA). Factor analyses differ from LCA and CA (and other person-centred techniques) in that they identify variables that are similar to each other, rather than observations. A study using data from Agincourt HDSS used 2 such techniques, principal component analysis and multiple correspondence analysis (MCA) on longitudinal data on household socio-economic status. The indices were created by pooling the data from all years even though some households would contribute data points than others. They also compared the PCA and MCA-generated indices with a much simpler to calculate index which did not require pooling and found the results to be similar (Kabudula et al., 2017). PCA has been used previously with Karonga HDSS data to generate SES indices (Kelly et al., 2018), however I was not able to do this in my analyses as household SES data has not been collected consistently throughout follow-up at the Karonga HDSS: the period used in the Kelly paper was shorter than in my own.

10.3.3. Repeated measures and missing data

In my literature review I noted that when faced with possible multiple events, many studies simply used the first event as the outcome, and excluded all data after that. In my thesis I attempted to use the data more fully, for example by specifically allowing events to be reversible in the sequence analysis so divorces and returns to school or the family home could be observed. Additionally in the migration analysis I used methods (multi-level modelling) which allowed for repeated competing events and possible clustering by household group. In the divorce analysis, I did use the first instance of divorce as the outcome, however, I was very careful to frame my research question and participant sample to explain this (i.e. specifically looking at divorce in the first 3 years of a first marriage).

I dealt with inconsistencies caused by repeated collection of the same data either by reducing the reports to one summary measure (i.e. with parental education status I took the highest report) or through data cleaning. For example, in the case of marital status an

impossible inconsistency would be a report of ‘never married’ following one of ‘divorced’. As I explained in the sequence analysis chapter, I created rules for such inconsistencies which were dependent on the reports before and after the inconsistency, and the length of time assigned to each status. I did not drop any records due to inconsistencies, even if they could not be resolved: this may have led to data issues as I highlighted in the sequence analysis chapter, i.e. where a period of ‘divorce’ may directly follow one of ‘never married’, without a period of ‘married’ in between. This was in contrast to another paper using HDSS data to examine adolescent transitions where a large number of records were dropped due to unresolvable inconsistencies (Del Fava et al., 2016). The reason I opted to keep all data was firstly one of practicality: I did not want to lose statistical power by dropping records; and secondly, I felt that dropping those with detectable inconsistencies would imply that the remaining data were fully correct. While it is tempting with the richness of HDSS data to use it to conduct detailed analysis or draw conclusions about specific individuals or small groups of individuals, it must be remembered that this is not what the system is designed for. Data are collected over a long period, by multiple, albeit trained, interviewers, often from proxies. Data may be reported differently from one round to another due to lack of knowledge from a proxy (i.e. they may not know that their wife’s niece who moved in recently had a short marriage), recall errors, different interpretation of events (i.e. one family member may not consider a short relationship a marriage, while another does), misunderstandings between the interviewer and interviewee, input errors by the interviewer, or through data entry. Once the data are collected, data manipulations may bring in further issues which may take the data used in an analysis further from the truth.

Using an example from my work: the marital status variables I derived came from annual reports of marital status, year of marriage/end of marriage for specific spouses and reports of age at first marriage (also usually reported annually). Precise dates of marriage are not recorded in the same way that dates of birth, death and migration are, so some assumptions are needed to be able to apply the marriage data to the continuous HDSS data. This may have introduced more noise and errors into the data. It is for this reason that I did not attempt to examine the order of different adolescent transitions, i.e. to identify if leaving school preceded marriage, as, unless the events happened with a long gap in between, any difference may be due to the data manipulation. A paper reflecting on an ethnographer’s experience comparing their detailed field data generated through multiple in-depth interviews and observations, with the data from the HDSS database really serves to highlight the issue I raised above (Reynolds, 2015). In the example given in this paper, even if the data is ‘correct’ the actual reason for the changes experienced by the person were much more nuanced that could be concluded from the HDSS data (Reynolds, 2015). The strength of the

HDSS data is in describing trends and patterns coming from the area, rather than trying to draw complex conclusions about individual's lives.

I chose to 'clean' data inconsistencies using my own rules, which are documented in my coding files. Manual attempts at data cleaning like this rely on the user to detect all data errors and to create appropriate rules to repair them. This may lead to some data errors being missed, and potentially the introduction of bias, i.e. by unconsciously choosing rules that favour the 'divorced' category being chosen for uncertain data. There have been developments regarding automated techniques to detect and repair data errors (Chu et al., 2016), which may be of some interest for use in HDSS data, especially as the datasets can get very large, however it was out of the scope of this thesis. For the sequence data I also dealt with some missing data in the same way as the inconsistencies. I did not attempt to repair or impute missing data anywhere else. In some cases, I left missingness as a separate category for some exposure variables in regression models, and in others I had to drop records with missing data, for example adolescents missing both parent IDs. A technique commonly used in research data to deal with missing data is multiple imputation. An imputation model is run a set number of times, producing that number of complete, but slightly different, datasets. The analytical model is then run on each dataset and the estimates produced from each model are combined, using a standard set of rules, to generate a single reportable set of estimates and precision measures (Kenward and Carpenter, 2007). I chose not to do this for my regression analyses as the level of missing data was quite low (for example, in the divorce analysis unknown father education accounted for only 48 of almost 4500 person years of follow-up), so I thought it would add unnecessary complexity. Methods are still being developed to use multiple imputation, or similar methods, for use in sequence analysis (Liao et al., 2022) so I was not able to consider using them for that analysis. Equally, the aim of that chapter was to examine issues relating to migration, so it would have added a layer of complexity for probably not much gain. That being said, I do recognise that methods to deal with missing data in longitudinal dataset such as HDSS are important and will be something I will continue to look at beyond this thesis.

10.3.4. Migrations

Considering the bias brought in by participants moving in and out of the area, and therefore the data, shaped all of my analyses. In the LCA household analysis, I opted to perform the LCA repeatedly on several cross-sectional snapshots rather than pooling all data points from all households across the time period because I was concerned that stable households

appearing multiple times in the dataset might change the results. The main thrust of the sequence analysis chapter was to access the usage of individual longitudinal data and the effect of migration on any conclusions. Migration itself was the outcome under consideration in the migration paper, but there were still some of the analyses where I needed to consider the effects of migration bias as some analyses were only possible with the subset of individuals who moved within the area. In the divorce paper I allowed out-migrants to contribute time to the analysis while they were present, however I was not able to include in-migrants so I tried to mitigate the effects of migration (that the people still in the analysis nearing the end of the analysis time might be different to the full group that started the analysis time) by only looking at a relatively short period of follow-up. The nature of HDSS data means that migration is something that will always need to be considered in almost all analyses beyond simple cross-sectional analyses. Careful use of the data may reduce the impact on results and conclusions, however it will never be possible to mitigate the effect entirely. As indicated in the literature review of other complex HDSS analyses in chapter 4, joint modelling is almost never used in HDSS data. This is a technique better suited to modelling time to event outcomes with longitudinal exposure compared to using Cox or Poisson regression with time-varying covariates, as the latter has been found to under-estimate the association (Sweeting and Thompson, 2011). It can also account for informative censoring so has the potential to be really useful for HDSS analyses. As it has mostly been developed regarding survival analyses in clinical trials, the main focus was on right censoring (participant drop-out), but there have also been developments to allow for left censoring (i.e. in-migration) (Crowther et al., 2016) and interval censoring (i.e. leaving and returning) (Lovblom et al., 2023).

10.4. Reproducibility of my work with other HDSS data

As I indicated in the introduction, while HDSS data are not as consistent and easily available as DHS data, there is general consistency in the way the standard HDSS data items are collected, and existing collaborative networks have harmonised and shared them (i.e. the INDEPTH network (http://www.indepth-network.org/) and the ALPHA network (Reniers et al., 2016)). This means that, with some additional work, it is possible to carry out similar or pooled analyses: there are many such examples of publications (i.e. Ginsburg et al. 2021; Marston et al. 2016). My analyses do harness aspects of the Karonga HDSS data, however, which are not standard; the detailed kinship links coupled with household level GPS data. Below I have summarised how other HDSSs deal with this sort of data to show how common such analyses might be possible. This information was gathered from published HDSS

protocols which are available for most HDSSs. These protocols tend to follow a similar format, however there are differences in how each report which data are collected and how: I often had to search a few different sections to find information on family linkages available, sometimes it was stated explicitly, sometimes mentioned briefly in a table. Prospective users of HDSS data for single, comparative and pooled analyses would be assisted greatly by a data atlas showing what data are available and for which years. An example of such a data atlas is one for longitudinal cohorts in the UK, which gives over-arching information on all the cohorts included and shows where there is overlap in the types of data collected (https://discovery.closer.ac.uk/). Obviously this requires dedicated funding and collaboration to set up and maintain.

Many HDSSs collect data suitable for studying the transition to adulthood: schooling and marital status are regularly collected, as are births with links to parent IDs. In fact, some HDSSs collect more detailed birth histories from in-migrant women than is done in Karonga, making studying that transition easier. Other markers of the transition to adulthood such as age at menarche and sexual debut may also be available in some HDSSs: these are available for some participants from the Karonga HDSS, however I did not use them in my analyses, as they were only collected for a short period of time. Indeed, aspects of the transition to adulthood has been studied using secondary data in Manicaland HDSS in Zimbabwe (Del Fava et al., 2016), uMkhanyakude HDSS in South Africa (Ardington et al., 2015) and Cuatro Santos HDSS in Nicaragua (Pérez et al., 2021); and using new qualitative data in Rakai HDSS in Uganda (Kreniske et al., 2019) and in Nairobi HDSS in Kenya (Pike et al., 2018).

10.4.1. Kinship data in other HDSSs

Karonga HDSS may be relatively unique with the extent of mother, father and spouse ID linkage which made my analyses of household structure and family nearby possible: the proportion of participants of all ages with mother and father ID links is high, although it is the highest for younger people. Most HDSSs record at least mother ID with birth registrations, allowing studies such as the pooled analysis of data from multiple HDSSs looking at child mortality around the timing of things like mother's death and sibling birth (Bocquier et al., 2021). Some additionally record the mother ID for some older children, i.e. Nahuche, Nigeria which links children under 5 with their mothers on vaccination data (Alabi et al., 2014). Father ID will also often be captured on birth registrations, and many HDSSs also regularly record the relationship of each person to the household head. For example, in the Mekong HDSS, only relationship to household head is reported and this has been used for analyses of household structure and child education (Heuveline and Hong, 2017). Other HDSSs which

seem to have maternal and paternal links for all ages similar to Karonga includes Mlomp HDSS, Senegal: the cohort profile specifically mentions the possibility of household structure and kinship studies (Pison et al., 2018) though there do not yet appear to be any published; Bandafassi and Niakhar HDSSs, also in Senegal, both have been used to generate lists of siblings for a study to validate a technique using sibling reports of mortality (Helleringer et al., 2014a, 2014b). In Bandafassi, the collection of detailed genealogies was begun early on, for use in genetics and anthropology studies and was continued (Pison et al., 2014), which is quite similar to the history of the Karonga data, where the genealogy work was begun to assist work on leprosy.

While the Karonga data is rich and valuable, it only indicates the presence or absence of relatives; we have no information on the actual social relationship between the people and whether their presence has a positive, negative, neutral, or mixed effect on the index. There are several examples from the literature which demonstrate that proximity may not be synonymous with support: an in-depth description of kinship networks in a matrilineal area of Zambia showed that support received from family members depends on sex, i.e. mother may expect support from sons, but not daughters, and from brothers but not sisters (Cliggett, 2001); a qualitative study in a rural area of Tanzania found a 50/50 split in whether living near kin was good (due to support received) or bad (due to conflicts (Hadley, 2004); and finally, father absence and single mother households are often seen as markers of vulnerability, but in South Africa it has been shown that the level of support given by the father was not dependent on whether he was living with the child (Madhavan et al., 2008). Additionally, full identification of all relatives relies on all participants being linked to their parents, spouses and children; if any data are missing then a relative might be present but undetectable. As indicated in the methodology chapter, availability of parent ID is higher for younger people, and certain age groups are more likely to than others to have more household members where the relationship is unknown. The value of HDSSs as platforms for kinship studies has been noted before, and in many sites, additional data capture has been carried out to supplement standard sources, often to combat the issue raised above. In Niakhar HDSS, Senegal, which captures parent link data for all ages as standard, similar to Karonga, additional data capture was carried out in 2014 and 2016 to obtained detailed data on social networks within and outside the HDSS (Delaunay et al., 2019). These data have been used in an analysis examining the effects of social ties on migration (Boujija et al., 2022). Matlab HDSS, Bangladesh records mother and father IDs for birth registrations, spouse IDs for marriages and relationship to household head as standard (Alam et al., 2017), and there appears to be data available from 1983-2001 which allows creation of networks of the compounds, based on maternal connections; this has been used to assess

the impact of social connectivity on diarrhoeal illness in children (Perez-Heydrich et al., 2013). Agincourt HDSS in South Africa collects relationship to household head as standard: this has been used in studies of household composition, for example in older people (Schatz et al., 2018). For a few waves of HDSS data they also collected a social connections database which consists of relationship data from the child's perspective. Data from 2002 from this database was used to assess household compositions and the association with school progress in children (Madhavan et al., 2017b). Also in Agincourt but later on, they collected even more in depth information on children's kinship networks (Madhavan et al., 2014). uMkhanyakude HDSS in South Africa collects relationship to household head as standard (Tanser et al., 2008), but in 2009 included a 'non-residents living arrangements survey' to help understand migrants and their relationships to source and destination households, whether children move with their parents and the characteristics of 'left behind' children (Bennett et al., 2015a, 2015b). And finally a study at Nairobi HDSS used a kinship tool to identify not only relatives outside of the household but an indicator of whether they provide support (Madhavan et al., 2017a).

10.4.2. Use of compound and other household membership definitions in other HDSSs

In Karonga, people do not tend to live in compounds, however often live close to, and have regular interactions with, close relatives. These household linkages are not captured in the data, however I used the household GPS data and kinship data to create these links myself. Many HDSSs do group households by compound, if this is most appropriate for the way the population live: Bandafassi, Senegal (Pison et al., 2014), Navrongo, Ghana (Oduro et al., 2012), KEMRI, Kenya (Odhiambo et al., 2012), Farafenni, Gambia (Jasseh et al., 2015), Niakhar, Senegal (Delaunay et al., 2013), Mbita, Kenya (Wanyua et al., 2013), Nahuche, Nigeria (Alabi et al., 2014), Nanoro, Burkina Faso (Derra et al., 2012), Kombewa, Kenya (Sifuna et al., 2014), uMkhanyakude, South Africa, (Hosegood et al., 2006) and Matlab, Bangladesh (Alam et al., 2017). However there does not seem to be many examples of examining the data by creating compound-level summary variables, or even using the compound ID as a clustering factor. Exceptions include a study with Nahuche HDSS data (Nigeria) which created compound-level variables by aggregating household information for analysis on child mortality (Alabi et al., 2017) and an analysis from Matlab HDSS (Bangladesh) creating networks of compound and the effect on childhood diarrhoeal disease (described in more detail in the above section on uses of kinship data in HDSSs) (Perez-Heydrich et al., 2013). Definitions of household or compound membership seem to be relatively similar across HDSS (eat from the same pot or kitchen and recognise the same

household head), but some have slightly different definitions, in response to the cultural conditions, which would allow for more nuanced analyses. For example, in South Africa, levels of circular labour migration are very high so Agincourt HDSS allows temporary migrants to be recorded (these are invisible in Karonga) (Kahn et al., 2012), and uMkhanyakude HDSS allows non-resident household members to be recorded (who must have spent at least 1 night in the area in the last 12 months), these non-residents may be resident in another bounded structure within the HDSS or outside of the area. Households can also have affiliate members, i.e. staff members, (Tanser et al., 2008).

10.5. Recommendations for further work

I have identified the following areas for further work using the Karonga or other HDSS data:

- Adolescence and transition to adulthood:
    - Investigation into the association between living arrangements and/or family with educational outcomes
    - Investigation into gender differences in living arrangements
    - Analysis of the association between socio-economic status and the transition to adulthood
    - Comparative and pooled analyses of adolescence and transition to adulthood using data from other HDSSs.
- Data issues:
    - Development and standardisation of automated documentation of complex data manipulations
    - Development and use of methods to deal with missing and/or inconsistent data from HDSSs
    - Encouragement of use of standard and more complex longitudinal techniques on HDSS data
    - Development of a documented and maintained data atlas for HDSS data

The core Karonga HDSS dataset described in the methodology chapter covers all ages so there are also many areas of family research that could be carried out: examples include changes of living arrangements in childhood, the affect of presence of family on childhood vaccine uptake, living arrangements in old age and the association with mortality.

10.6. Conclusions

In this thesis I have demonstrated the flexibility and utility of HDSS data by using a range of different sophisticated data manipulation and statistical techniques to answer specific research questions related to adolescence and the transition to adulthood. My findings show that family, beyond parents, is important for many adolescents in this area of rural northern Malawi, affecting aspects of the transition to adulthood, including migration and schooling. The transition to adulthood tends to follow quite traditional pathways, though marital age may be increasing for some sections of the population, and divorce is common. There does not yet seem to be much evidence of 'new' types of transition to adulthood, for example including periods of migration to urban areas for education or work.

I have also made methodological contributions to the literature, and increased awareness and visibility of HDSS by assessing the value of these techniques for use with this data source. I have identified areas of development and collaboration which would substantively improve use of, and access to HDSS data, namely, data processing documentation standards, and a data atlas of HDSS data. These infrastructure suggestions can only function with the experience of local HDSS data producers, with appropriate expertise in longitudinal data processing and analyses. These collaborative works would require substantial investment by funders, both for developing infrastructure and for training and supporting HDSS data specialists. The expertise of these professionals should be recognised within their own organisations, and the wider research community, in terms of credit on grant applications and research outputs.

11. References

Adjiwanou, V., Boco, G.A., Yaya, S., 2021. Stepfather families and children's schooling in sub-Saharan Africa: A cross-national study. Demogr. Res. 44, 627–670.

Agarwal, D., Dhotre, D., Patil, R., Shouche, Y., Juvekar, S., Salvi, S., 2017. Potential of Health and Demographic Surveillance System in Asthma and Chronic Obstructive Pulmonary Disease Microbiome Research. Front. Public Heal. 5, 262279.

Agergaard, J., 1999. The household as a unit of analysis: Reflections from migration research in Nepal. Geogr. Tidsskr. 99, 101–111.

Akinyemi, J.O., Chisumpa, V.H., Odimegwu, C.O., 2016. Household structure, maternal characteristics and childhood mortality in rural sub-Saharan Africa. Rural Remote Health 16, 3737.

Alabi, O., Doctor, H. V, Jumare, A., Sahabi, N., Abdulwahab, A., Findley, S.E., Abubakar, S.D., 2014. Health & Demographic Surveillance System Profile: The Nahuche Health and Demographic Surveillance System, Northern Nigeria (Nahuche HDSS). Int. J. Epidemiol. 43, 1770–1780.

Alabi, O., Oyedokun, O.A., Doctor, H. V., Adedini, S.A., 2017. Determinants of under-five mortality clustering in a health and demographic surveillance system in Zamfara State, northern Nigeria. African Popul. Stud. 31.

Alam, N., Ali, T., Razzaque, A., Rahman, M., Zahirul Haq, M., Saha, S.K., Ahmed, A., Sarder, A., Moinuddin Haider, M., Yunus, M., Nahar, Q., Kim Streatfield, P., 2017. Health and Demographic Surveillance System (HDSS) in Matlab, Bangladesh. Int. J. Epidemiol. 46, 809–816.

Alter, G., Donakowski, D., Gager, J., Heus, P., Hunter, C., Ionescu, S., Iverson, J., Jagadish, H.V., Lagoze, C., Lyle, J., Mueller, A., Revheim, S., Richardson, M.A., Ørnulf, R., Seelam, K., Smith, D., Smith, T., Song, J., Vaidya, Y.J., Voldsater, O., 2020. Provenance metadata for statistical data: An introduction to Structured Data Transformation Language (SDTL). IASSIST Q. 44, 1–26.

Anekwe, T.D., Newell, M.L., Tanser, F., Pillay, D., Bärnighausen, T., 2015. The causal effect of childhood measles vaccination on educational attainment: A mother fixed-effects study in rural South Africa. Vaccine 33, 5020–5026.

Anglewicz, P., Reniers, G., 2014. HIV status, gender, and marriage dynamics among adults in Rural Malawi. Stud. Fam. Plann. 45, 415–28.

Angotti, N., Houle, B., Schatz, E., Mojola, S., 2018. Classifying and Contextualizing Sexual Practices across the Life Course: Implications in Later Life. In: PAA Conference.

Ansell, N., Hajdu, F., van Blerk, L., Robson, E., 2018. "My happiest time" or "my saddest

time”? The spatial and generational construction of marriage among youth in rural Malawi and Lesotho. Trans. Inst. Br. Geogr. 43, 184–199.

Ardington, C., Bärnighausen, T., Case, A., Menendez, A., 2014. The economic consequences of AIDS mortality in South Africa. J. Dev. Econ. 111, 48–60.

Ardington, C., Menendez, A., Mutevedzi, T., 2015. Early childbearing, human capital attainment, and mortality risk: Evidence from a longitudinal demographic surveillance area in rural KwaZulu-Natal, South Africa. Econ. Dev. Cult. Change 63, 281–317.

Arthur, S., Bangha, M., Sankoh, O., 2013. Review of contributions from HDSSs to research in sexual and reproductive health in low- and middle-income countries. Trop. Med. Int. Heal. 18, 1463–1487.

Asiki, G., Newton, R., Marions, L., Kamali, A., Smedman, L., 2019. The effect of childhood stunting and wasting on adolescent cardiovascular diseases risk and educational achievement in rural Uganda: a retrospective cohort study. Glob. Health Action 12, 1626184.

Bagayoko, M., Akeyo, D., Kadengye, D.T., Iddi, S., 2020. Understanding wealth transitions among households in urban slums of Nairobi: A multi-state transition modelling approach. Glob. Epidemiol. 2, 100037.

Bairagi, R., Becker, S., Kantner, A., Allen, K.B., Datta, A., Purvis, K., 1997. An evaluation of the 1993-94 Bangladesh Demographic and Health Survey within the Matlab area.

Baruwa, O.J., Amoateng, A.Y., Biney, E., 2019. Socio-demographic changes in age at first marriage in Malawi: Evidence from Malawi Demographic and Health Survey data, 1992-2016. J. Biosoc. Sci. 52, 832–845.

Beaman, L., Dillon, A., 2012. Do household definitions matter in survey design? Results from a randomized survey experiment in Mali. J. Dev. Econ. 98, 124–135.

Becher, H., Müller, O., Dambach, P., Gabrysch, S., Niamba, L., Sankoh, O., Simboro, S., Schoeps, A., Stieglbauer, G., Yé, Y., Sié, A., 2016. Decreasing child mortality, spatial clustering and decreasing disparity in North-Western Burkina Faso. Trop. Med. Int. Heal. 21, 546–555.

Beegle, K., Poulin, M., 2013. Migration and the Transition to Adulthood in Contemporary Malawi. Ann. Am. Acad. Pol. Soc. Sci. 648, 38–51.

Beegle, K., Poulin, M., 2017. Marriage Transitions in Malawi Panel Data. Stud. Fam. Plann. 48, 391–396.

Beguy, D., Kabiru, C.W., Zulu, E.M., Ezeh, A.C., 2011. Timing and sequencing of events marking the transition to adulthood in two informal settlements in Nairobi, Kenya. J. Urban Health 88 Suppl 2, S318-40.

Bellis, M.A., Downing, J., Ashton, J.R., 2006. Adults at 12? Trends in puberty and their public health consequences. J. Epidemiol. Community Health.

Bennett, R., Hosegood, V., Newell, M., McGrath, N., 2015a. An Approach to Measuring Dispersed Families with a Particular Focus on Children 'Left Behind' by Migrant Parents: Findings from Rural South Africa. Popul. Space Place 21, 322–334.

Bennett, R., Hosegood, V., Newell, M., McGrath, N., 2015b. Understanding Family Migration in Rural South Africa: Exploring Children's Inclusion in the Destination Households of Migrant Parents. Popul. Space Place 21, 310–321.

Bennett, R., Waterhouse, P., 2018. Work and family transitions and the self-rated health of young women in South Africa. Soc. Sci. Med. 203, 9–18.

Bertrand-Dansereau, A., Clark, S., 2016. Pragmatic tradition or romantic aspiration? The causes of impulsive marriage and early divorce among women in rural Malawi. Demogr. Res. 35, 47–80.

Biddecom, A., Bakilana, A., 2003. Transitions into sex, parenthood and unions among adolescents and young adults in South Africa. Natl. Inst. Child Heal. Hum. Dev.

Bignami-Van Assche, S., Boulet, V., Simard, C.-O., 2021. A New Methodological Approach to Study Household Structure From Census and Survey Data. Sociol. Methods Res. 004912412098619.

Birdthistle, I., Tanton, C., Tomita, A., de Graaf, K., Schaffnit, S.B., Tanser, F., Slaymaker, E., 2019. Recent levels and trends in HIV incidence rates among adolescent girls and young women in ten high-prevalence African countries: a systematic review and meta-analysis. Lancet Glob. Heal. 7, e1521–e1540.

Bishai, D., Razzaque, A., Christiansen, S., Mustafa, A.H.M.G., Hindin, M., 2015. Selection Bias in the Link Between Child Wantedness and Child Survival: Theory and Data From Matlab, Bangladesh. Demography 52, 61–82.

Blakemore, S.J., Choudhury, S., 2006. Development of the adolescent brain: Implications for executive function and social cognition. J. Child Psychol. Psychiatry Allied Discip.

Bocquier, P., 2016. Migration Analysis Using Demographic Surveys and Surveillance Systems. In: International Handbook of Migration and Population Distribution. Springer, Dordrecht, pp. 205–223.

Bocquier, P., Ginsburg, C., Collinson, M.A., 2019. A training manual for event history analysis using longitudinal data. BMC Res. Notes 12, 506.

Bocquier, P., Ginsburg, C., Herbst, K., Sankoh, O., Collinson, M.A., 2017a. A training manual for event history data management using Health and Demographic Surveillance System data. BMC Res. Notes 10, 224.

Bocquier, P., Ginsburg, C., Menashe-Oren, A., Compaoré, Y., Collinson, M., 2021. The crucial role of mothers and siblings in child survival: Evidence from 29 health and demographic surveillance systems in sub-saharan africa. Demography 58, 1687–1713.

Bocquier, P., Sankoh, O., Byass, P., 2017b. Are health and demographic surveillance

system estimates sufficiently generalisable? Glob. Health Action 10.

Boujija, Y., Bignami, S., Delaunay, V., Sandberg, J., 2022. Who Matters Most? Migrant Networks, Tie Strength, and First Rural–Urban Migration to Dakar. Demography 59, 1683–1711.

Breton, E., 2019. Modernization and Household Composition in India, 1983–2009. Popul. Dev. Rev. 45, 739–766.

Bronte-Tinkew, J., Dejong, G., 2004. Children's nutrition in Jamaica: Do household structure and household economic resources matter? Soc. Sci. Med. 58, 499–514.

Brzinsky-Fay, C., 2014. Graphical Representation of Transitions and Sequences. In: Life Course Research and Social Policies. Springer Science and Business Media B.V., pp. 265–284.

Camlin, C.S., Snow, R.C., Hosegood, V., 2014. Gendered Patterns of Migration in Rural South Africa. Popul. Space Place 20, 528–551.

Chae, S., 2016. Forgotten marriages? Measuring the reliability of marriage histories. Demogr. Res. 34, 525–562.

Chae, S., Hayford, S.R., Agadjanian, V., 2016. Father's Migration and Leaving the Parental Home in Rural Mozambique. J. Marriage Fam. 78, 1047–1062.

Chalasani, S., Mensch, B.S., Hewett, P.C., 2013. Migration among adolescents from rural Malawi. In: Annual Meeting of the Population Association of America.

Chandramohan, D., Shibuya, K., Setel, P., Cairncross, S., Lopez, A.D., Murray, C.J.L., Żaba, B., Snow, R.W., Binka, F., 2008. Should Data from Demographic Surveillance Systems Be Made More Widely Available to Researchers? PLoS Med. 5, e57.

Chang, A.Y., Haber, N., Bärnighausen, T., Herbst, K., Gareta, D., Pillay, D., Salomon, J.A., 2018. Improving the Validity of Mathematical Models for HIV Elimination by Incorporating Empirical Estimates of Progression Through the HIV Treatment Cascade. JAIDS J. Acquir. Immune Defic. Syndr. 79, 596–604.

Charlton, C., Rasbash, J., Browne, W.J., Healy, M., Cameron, B., 2020. MLwiN Version 3.05. Centre for Multilevel Modelling, University of Bristol.

Chawinga, W.D., Zinn, S., 2019. Global perspectives of research data sharing: A systematic literature review. Libr. Inf. Sci. Res.

Chimango, L.J., 1977. Woman without "ankhoswe" in malawi: A discussion of the legal position of women who enter into informal marital relations. J. Leg. Plur. Unoff. Law 9, 54–61.

Cho, G.J., Shin, J.-H., Yi, K.W., Park, H.T., Kim, T., Hur, J.Y., Kim, S.H., 2012. Adolescent pregnancy is associated with osteoporosis in postmenopausal women. Menopause 19, 456–460.

Chu, X., Ilyas, I.F., Krishnan, S., Wang, J., 2016. Data cleaning: Overview and emerging

challenges. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Association for Computing Machinery, pp. 2201–2206.

Clark, S., Brauner-Otto, S., 2015. Divorce in sub-Saharan Africa: Are Unions Becoming Less Stable? Popul. Dev. Rev. 41, 583–605.

Clark, S., Cotton, C., 2013. Transitions to adulthood in urban Kenya: A focus on adolescent migrants. Demogr. Res. 28.

Clark, S.J., Kahn, K., Houle, B., Arteche, A., Collinson, M.A., Tollman, S.M., Stein, A., 2013. Young Children's Probability of Dying Before and After Their Mother's Death: A Rural South African Population-Based Surveillance Study. PLoS Med. 10, e1001409.

Cleland, J., 2010. Demographic Data Collection in Less Developed Countries 1946–1996. Popul. Stud. (NY). 50, 433–450.

Cliggett, L., 2000. Social components of migration: Experiences from Southern Province, Zambia. Hum. Organ. 59, 125–135.

Cliggett, L., 2001. Survival strategies of the elderly in Gwembe Valley, Zambia: Gender, residence and kin networks. J. Cross. Cult. Gerontol. 16, 309–332.

Coall, D.A., Tickner, M., McAllister, L.S., Sheppard, P., 2016. Developmental influences on fertility decisions by women: An evolutionary perspective. Philos. Trans. R. Soc. B Biol. Sci.

Colombe, S., Beard, J., Mtenga, B., Lutonja, P., Mngara, J., De Dood, C.J., Van Dam, G.J., Corstjens, P.L.A.M., Kalluvya, S., Urassa, M., Todd, J., Downs, J.A., 2019. HIV-seroconversion among HIV-1 serodiscordant married couples in Tanzania: A cohort study. BMC Infect. Dis. 19, 518.

Corsi, D.J., Neuman, M., Finlay, J.E., Subramanian, S., 2012. Demographic and health surveys: a profile. Int. J. Epidemiol. 41, 1602–1613.

Crampin, A.C., Dube, A., Mboma, S., Price, A., Chihana, M., Jahn, A., Baschieri, A., Molesworth, A., Mwaiyeghele, E., Branson, K., Floyd, S., McGrath, N., Fine, P.E.M., French, N., Glynn, J.R., Zaba, B., 2012. Profile: the Karonga Health and Demographic Surveillance System. Int. J. Epidemiol. 41, 676–85.

Crowther, M.J., Andersson, T.M.-., Lambert, P.C., Abrams, K.R., Humphreys, K., 2016. Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. Stat. Med. 35, 1193–1209.

Daniel, A., 2017. Malawi Amends Constitution to Remove Child Marriage Loophole | Human Rights Watch [WWW Document]. Hum. Rights Watch. URL https://www.hrw.org/news/2017/02/23/malawi-amends-constitution-remove-child-marriage-loophole (accessed 2.12.20).

Day, C., Evans, R., 2015. Caring Responsibilities, Change and Transitions in Young People's Family Lives in Zambia. J. Comp. Fam. Stud. XLVI, 137–152.

Day, F.R., Elks, C.E., Murray, A., Ong, K.K., Perry, J.R.B., 2015. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the UK Biobank study. Sci. Rep. 5, 11208.

de Walque, D., Kline, R., 2012. The Association Between Remarriage and HIV Infection in 13 Sub-Saharan African Countries. Stud. Fam. Plann. 43, 1–10.

De Wet, N., 2012. Parent absenteeism and adolescent work in South Africa: An analysis of the levels and determinants of adolescents who work 10 or more hours a week. Etude la Popul. Africaine 27, 70–78.

Del Fava, E., Piccarreta, R., Gregson, S., Melegaro, A., 2016. Transition to Parenthood and HIV Infection in Rural Zimbabwe. PLoS One 11, e0163730.

Delaunay, V., Douillot, L., Diallo, A., Dione, D., Trape, J.-F., Medianikov, O., Raoult, D., Sokhna, C., 2013. Profile: The Niakhar Health and Demographic Surveillance System. Int. J. Epidemiol. 42, 1002–1011.

Delaunay, V., Douillot, L., Rytina, S., Boujija, Y., Bignami, S., Ba Gning, S., Sokhna, C., Belaid, L., Fotouhi, B., Senghor, A., Sandberg, J., 2019. The Niakhar Social Networks and Health Project. MethodsX 6, 1360–1369.

Delprato, M., Akyeampong, K., Dunne, M., 2017. Intergenerational Education Effects of Early Marriage in Sub-Saharan Africa. World Dev. 91, 173–192.

Delprato, M., Akyeampong, K., Sabates, R., Hernandez-Fernandez, J., 2015. On the impact of early marriage on schooling outcomes in Sub-Saharan Africa and South West Asia. Int. J. Educ. Dev. 44, 42–55.

Deribew, A., Ojal, J., Karia, B., Bauni, E., Oteinde, M., 2016. Under-five mortality rate variation between the Health and Demographic Surveillance System (HDSS) and Demographic and Health Survey (DHS) approaches. BMC Public Health 16, 1–7.

Derra, K., Rouamba, E., Kazienga, A., Ouedraogo, S., Tahita, M.C., Sorgho, H., Valea, I., Tinto, H., 2012. Profile: Nanoro Health and Demographic Surveillance System. Int. J. Epidemiol. 41, 1293–1301.

Dobra, A., Bärnighausen, T., Vandormael, A., Tanser, F., 2017. Space-time migration patterns and risk of HIV acquisition in rural South Africa. AIDS 31, 137–145.

Dobra, A., Bärnighausen, T., Vandormael, A., Tanser, F., 2019. A method for statistical analysis of repeated residential movements to link human mobility and HIV acquisition. PLoS One 14, e0217284.

Duc, D.M., Vui, L.T., Son, H.N., Minh, H. Van, 2016. Smoking Initiation and Cessation among Youths in Vietnam: A Longitudinal Study Using the Chi Linh Demographic—Epidemiological Surveillance System (CHILILAB DESS). AIMS Public Heal. 4, 1–18.

Dzomba, A., Tomita, A., Govender, K., Tanser, F., 2019. Effects of Migration on Risky Sexual Behavior and HIV Acquisition in South Africa: A Systematic Review and Meta-

analysis, 2000–2017. AIDS Behav.

Edson Utazi, C., Sahu, S.K., Atkinson, P.M., Tejedor-Garavito, N., Lloyd, C.T., Tatem, A.J., 2018. Geographic coverage of demographic surveillance systems for characterising the drivers of childhood mortality in sub-Saharan Africa. BMJ Glob. Heal. 3, 611.

Etoori, D., Rice, B., Reniers, G., Gomez-Olive, F.X., Renju, J., Kabudula, C.W., Wringe, A., 2021. Patterns of engagement in HIV care during pregnancy and breastfeeding: findings from a cohort study in North-Eastern South Africa. BMC Public Health 21, 1–12.

Eyre, R.W., House, T., Xavier Gómez-Olivé, F., Griffiths, F.E., 2021. Bayesian belief network modelling of household food security in rural South Africa. BMC Public Health 21, 1–16.

Fabic, M.S., Choi, Y., Bird, S., 2012. Systematic reviews A systematic review of Demographic and Health Surveys: data availability and utilization for research. Bull World Heal. Organ 90, 604–612.

Farrell, K., 2017. The Rapid Urban Growth Triad: A New Conceptual Framework for Examining the Urban Transition in Developing Countries. Sustainability 9, 1407.

Finlay, J.E., Moucheraud, C., Goshev, S., Levira, F., Mrema, S., Canning, D., Masanja, H., Yamin, A.E., 2015. The Effects of Maternal Mortality on Infant and Child Survival in Rural Tanzania: A Cohort Study. Matern. Child Health J. 19, 2393–2402.

Flinn, M. V, Quinlan, R.J., Coe, K., Ward, C. V, 2007. 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.

Floyd, S., Molesworth, A., Dube, A., Crampin, A.C., Houben, R., Chihana, M., Price, A., Kayuni, N., Saul, J., French, N., Glynn, J.R., 2013. Underestimation of HIV prevalence in surveys when some people already know their status, and ways to reduce the bias. AIDS 27, 233–242.

Ford, K., Hosegood, V., 2005. AIDS mortality and the mobility of children in KwaZulu Natal, South Africa. Demography 42, 757–68.

Fotso, J.C., Cleland, J., Mberu, B., Mutua, M., Elungata, P., 2013. Birth spacing and child mortality: An analysis of prospective data from the nairobi urban health and demographic surveillance system. J. Biosoc. Sci. 45, 779–798.

Fottrell, E., Enquselassie, F., Byass, P., 2010. The distribution and effects of child mortality risk factors in Ethiopia: A comparison of estimates from DSS and DHS. Ethiop. J. Heal. Dev. 23, 163–168.

Furnas, H.E., 2016. Capturing Complexities of Relationship-Level Family Planning Trajectories in Malawi. Stud. Fam. Plann. 47, 205–221.

Gabadinho, A., Ritschard, G., Muller, N., Studer, M., 2011. Analyzing and visualizing state

sequences in R with TraMineR. J. Stat. Softw. 40, 1–37.

Gage, A.J., Sommerfelt, A.E., Piani, A.L., 1997. Household structure and childhood immunization in Niger and Nigeria. Demography 34, 295–309.

Gandrud, C., Allaire, J., Russell, K., Yetman, C., 2017. D3 JavaScript Network Graphs from R [WWW Document]. URL http://christophergandrud.github.io/networkD3/

Gansaonré, R.J., Moore, L., Bleau, L., Kobiané, J., Haddad, S., 2022. Stunting, age at school entry and academic performance in developing countries: A systematic review and meta-analysis. Acta Paediatr. 111, 1853–1861.

Gaydosh, L., 2015. Childhood Risk of Parental Absence in Tanzania. Demography 52, 1121–1146.

Gaydosh, L., 2017. Beyond Orphanhood: Parental Nonresidence and Child Well-being in Tanzania. J. Marriage Fam. 79, 1369–1387.

Ghafur, T., Islam, M.M., Alam, N., Hasan, M.S., 2020. Health and Demographic Surveillance System Sites: Reflections on Global Health Research Ethics. J. Popul. Soc. Stud. 28, 265–275.

Ginsburg, C., Bocquier, P., Béguy, D., Afolabi, S., Augusto, O., Derra, K., Herbst, K., Lankoande, B., Odhiambo, F., Otiende, M., Soura, A., Wamukoya, M., Zabré, P., White, M.J., Collinson, M.A., 2016. Healthy or unhealthy migrants? Identifying internal migration effects on mortality in Africa using health and demographic surveillance systems of the INDEPTH network. Soc. Sci. Med. 164, 59–73.

Ginsburg, C., Bocquier, P., Menashe-Oren, A., Collinson, M.A., 2021. Migrant health penalty: evidence of higher mortality risk among internal migrants in sub-Saharan Africa. Glob. Health Action 14.

Ginsburg, C., Norris, S.A., Richter, L.M., Coplan, D.B., 2009. Patterns of residential mobility amongst children in Greater Johannesburg-Soweto, South Africa: Observations from the birth to Twenty cohort. Urban Forum 20, 397–413.

Glynn, J.R., Kayuni, N., Floyd, S., Banda, E., Francis-Chizororo, M., Tanton, C., Molesworth, A., Hemmings, J., Crampin, A.C., French, N., 2010. Age at menarche, schooling, and sexual debut in northern Malawi. PLoS One 5, e15334.

Glynn, J.R., Sunny, B.S., DeStavola, B., Dube, A., Chihana, M., Price, A.J., Crampin, A.C., 2018. Early school failure predicts teenage pregnancy and marriage: A large population-based cohort study in northern Malawi. PLoS One 13, e0196041.

Goldberg, R.E., 2013. Family Instability and Pathways to Adulthood in Urban South Africa. Popul. Dev. Rev. 39, 231–256.

Goode, W., 1993. World Changes in Divorce Patterns. Yale University Press, New Haven.

Grant, M.J., 2012. Girls' schooling and the perceived threat of adolescent sexual activity in rural Malawi. Cult. Heal. Sex. 14, 73–86.

Grant, M.J., Pike, I., 2019. Divorce, living arrangements, and material well-being during the transition to adulthood in rural Malawi. Popul. Stud. (NY). 73, 261–275.

Grant, M.J., Soler-Hampejsek, E., 2014. HIV Risk Perceptions, the Transition to Marriage, and Divorce in Southern Malawi. Stud. Fam. Plann. 45, 315–337.

Grant, M.J., Yeatman, S., 2014. The impact of family transitions on child fostering in rural Malawi. Demography 51, 205–28.

Graybill, L.A., Kasaro, M., Freeborn, K., Walker, J.S., Poole, C., Powers, K.A., Mollan, K.R., Rosenberg, N.E., Vermund, S.H., Mutale, W., Chi, B.H., 2020. Incident HIV among pregnant and breast-feeding women in sub-Saharan Africa: A systematic review and meta-analysis. AIDS 34, 761–776.

Grieger, L., Williamson, A., Leibbrandt, M., Levinsohn, J., 2013. Moving out and moving in: Evidence of short-term household change in South Africa from the National Income Dynamics Study. A Southern Africa Labour and Development Research Unit Working Paper Number 106. SALDRU, University of Cape Town, Cape Town.

Guirkinger, C., Gross, J., Platteau, J.P., 2021. Are women emancipating? Evidence from marriage, divorce and remarriage in Rural Northern Burkina Faso☆. World Dev. 146, 105512.

Haber, N., Tanser, F., Bor, J., Naidu, K., Mutevedzi, T., Herbst, K., Porter, K., Pillay, D., Bärnighausen, T., 2017. From HIV infection to therapeutic response: a population-based longitudinal HIV cascade-of-care study in KwaZulu-Natal, South Africa. Lancet HIV 4, e223–e230.

Hadley, C., 2004. The costs and benefits of kin: Kin networks and children's health among the Pimbwe of Tanzania. Hum. Nat. 15, 377–395.

Hampshire, K., Porter, G., Agblorti, S., Robson, E., Munthali, A., Abane, A., 2015. Context matters: Fostering, orphanhood and schooling in sub-Saharan Africa. J. Biosoc. Sci. 47, 141–164.

Harling, G., Newell, M.L., Tanser, F., Kawachi, I., Subramanian, S. V., Bärnighausen, T., 2014. Do age-disparate relationships drive HIV incidence in young women? Evidence from a population cohort in Rural KwaZulu-Natal, South Africa. J. Acquir. Immune Defic. Syndr. 66, 443–451.

Harper, S., Seekings, J., 2010. Claims on and Obligations to Kin in Cape Town, South Africa. University of Cape Town.

Hashim, I., 2007. Independent Child Migration and Education in Ghana. Dev. Change 38, 911–931.

Hedges, S., Sear, R., Todd, J., Urassa, M., Lawson, D., 2019. Earning their keep? Fostering, children's education, and work in north-western Tanzania. Demogr. Res. 41, 263–292.

Helleringer, S., Pison, G., Kanté, A.M., Duthé, G., Andro, A., 2014a. Reporting Errors in Siblings' Survival Histories and Their Impact on Adult Mortality Estimates: Results From a Record Linkage Study in Senegal. Demography 51, 387–411.

Helleringer, S., Pison, G., Masquelier, B., Kanté, A.M., Douillot, L., Duthé, G., Sokhna, C., Delaunay, V., 2014b. Improving the Quality of Adult Mortality Data Collected in Demographic Surveys: Validation Study of a New Siblings' Survival Questionnaire in Niakhar, Senegal. PLoS Med. 11, e1001652.

Hemmings, J., 2007. Infertility and Women's Life Courses in Northern Malawi. London School of Hygiene and Tropical Medicine.

Herbst, K., Juvekar, S., Bhattacharjee, T., Bangha, M., Patharia, N., Tei, T., Gilbert, B., Sankoh, O., 2015a. The INDEPTH Data Repository: An International Resource for Longitudinal Population and Health Data from Health and Demographic Surveillance Systems. J. Empir. Res. Hum. Res. Ethics.

Herbst, K., Juvekar, S., Jasseh, M., Berhane, Y., Chuc, N.T.K., Seeley, J., Sankoh, O., Clark, S.J., Collinson, M.A., 2021. Health and demographic surveillance systems in low- and middle-income countries: history, state of the art and future prospects. Glob. Health Action.

Herbst, K., Law, M., Geldsetzer, P., Tanser, F., Harling, G., Bärnighausen, T., 2015b. Innovations in health and demographic surveillance systems to establish the causal impacts of HIV policies. Curr. Opin. HIV AIDS.

Herrera-Almanza, C., Sahn, D.E., 2020. Childhood determinants of internal youth migration in Senegal. Demogr. Res. 43, 1335–1366.

Hertrich, V., Feuillet, P., Samuel, O., Doumbia Gakou, A., Dasré, A., 2020. Can we study the family environment through census data? A comparison of households, dwellings, and domestic units in rural Mali. Popul. Stud. (NY). 74, 119–138.

Hertrich, V., Lesclingand, M., 2012. Adolescent migration and the 1990s nuptiality transition in Mali. Popul. Stud. (NY). 66, 147–166.

Hertrich, V., Lesclingand, M., 2013. Adolescent Migration in Rural Africa as a Challenge to Gender and Intergenerational Relationships. Ann. Am. Acad. Pol. Soc. Sci. 648, 175–188.

Heuveline, P., Hong, S., 2017. Household structure and child education in Cambodia. Int. J. Popul. Stud. 3, 1.

Hewett, P.C., Mensch, B.S., 2019. Malawi Schooling and Adolescent Survey (MSAS) [WWW Document]. Harvard Dataverse. URL https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/V4C81G (accessed 2.27.13).

Hinga, A.N., Molyneux, S., Marsh, V., 2021. Towards an appropriate ethics framework for

Health and Demographic Surveillance Systems (HDSS): Learning from issues faced in diverse HDSS in sub-Saharan Africa. BMJ Glob. Heal. 6, 4008.

Hosegood, V., Benzler, J., Solarsh, G.C., 2005. Population mobility and household dynamics in rural South Africa: implications for demographic and health research. South. African J. Demogr.

Hosegood, V., Timaeus, I.M., Timaeus, I.M., 2006. Household Composition and Dynamics in KwaZulu Natal, South Africa: Mirroring Social Reality in Longitudinal Data Collection. In: African Households: Censuses and Surveys. Routledge, pp. 98–117.

Houle, B., Kabudula, C.W., Stein, A., Gareta, D., Herbst, K., Clark, S.J., 2021. Linking the timing of a mother's and child's death: Comparative evidence from two rural South African population-based surveillance studies, 2000–2015. PLoS One 16, e0246671.

Houle, B., Stein, A., Kahn, K., Madhavan, S., Collinson, M., Tollman, S.M., Clark, S.J., 2013. Household context and child mortality in rural South Africa: the effects of birth spacing, shared mortality, household composition and socio-economic status. Int. J. Epidemiol. 42, 1444–54.

Hownana, A., 2012. "Desenrascar a Vida": Youth Employment and Transitions to Adulthood. In: Third Conference of IESE. Maputo.

Huffman, C., Regules-García, R., Vargas-Chanes, D., 2019a. Living arrangement dynamics of older adults in Mexico: Latent class analysis in an accelerated longitudinal design. Demogr. Res. 41, 1401–1436.

Huffman, C., Villagómez-Ornelas, P., Vargas Chanes, D., 2019b. Family arrangements and savings in Mexico: a latent class approach.

Hunleth, J., Jacob, R., Cole, S., Bond, V., James, A., 2015. School holidays: examining childhood, gender norms, and kinship in children's shorter-term residential mobility in urban Zambia. Child. Geogr. 13, 501–517.

Hunter, L.M., Talbot, C., Twine, W., McGlinchy, J., Kabudula, C.W., Ohene-Kwofie, D., 2021. Working toward effective anonymization for surveillance data: innovation at South Africa's Agincourt Health and Socio-Demographic Surveillance Site. Popul. Environ. 42, 445–476.

Hwang, W., Jung, E., 2021. Helicopter Parenting Versus Autonomy Supportive Parenting? A Latent Class Analysis of Parenting Among Emerging Adults and Their Psychological and Relational Well-Being. Emerg. Adulthood.

Ijadunola, M.Y., Obiyan, M.O., Odia, B.., Ogun, O.O., Fajinmi, O.., Uzezi, P.., Abiola, A.., 2017. Assessing the Challenges of Schooling among Adolescents in Skipped Generation Households in Ile-Ife, Nigeria, Journal of Community Medicine and Primary Health Care.

Jasseh, M., Gomez, P., Greenwood, B.M., Howie, S.R., Scott, S., Snell, P.C., Bojang, K.,

Cham, M., Corrah, T., D'Alessandro, U., 2015. Health & Demographic Surveillance System Profile: Farafenni Health and Demographic Surveillance System in The Gambia. Int. J. Epidemiol. 44, 837–847.

Jasseh, M., Rerimoi, A.J., Reniers, G., Timæus, I.M., 2022. Assessment of the consistency of health and demographic surveillance and household survey data: A demonstration at two HDSS sites in The Gambia. PLoS One 17, e0271464.

Jeong, J., 2021. Determinants and Consequences of Adolescent Fatherhood: A Longitudinal Study in Ethiopia, India, Peru, and Vietnam. J. Adolesc. Heal. 68, 906–913.

Jia, P., Sankoh, O., Tatem, A.J., 2015. Mapping the environmental and socioeconomic coverage of the INDEPTH international health and demographic surveillance system network. Heal. Place 36, 88–96.

Juárez, F., Gayet, C., 2014. Transitions to Adulthood in Developing Countries. Annu. Rev. Sociol. 40, 521–538.

Kabudula, C.W., Houle, B., Collinson, M.A., Kahn, K., Tollman, S., Clark, S., 2017. Assessing Changes in Household Socioeconomic Status in Rural South Africa, 2001–2013: A Distributional Analysis Using Household Asset Indicators. Soc. Indic. Res. 133, 1047–1073.

Kahn, K., Collinson, M.A., Gomez-Olive, F.X., Mokoena, O., Twine, R., Mee, P., Afolabi, S.A., Clark, B.D., Kabudula, C.W., Khosa, A., Khoza, S., Shabangu, M.G., Silaule, B., Tibane, J.B., Wagner, R.G., Garenne, M.L., Clark, S.J., Tollman, S.M., 2012. Profile: Agincourt Health and Socio-demographic Surveillance System. Int. J. Epidemiol. 41, 988–1001.

Källestål, C., Blandón, E.Z., Peña, R., Peréz, W., Contreras, M., Persson, L.-Å., Sysoev, O., Selling, K.E., 2020. Assessing the Multiple Dimensions of Poverty. Data Mining Approaches to the 2004–14 Health and Demographic Surveillance System in Cuatro Santos, Nicaragua. Front. Public Heal. 7, 409.

Kanjala, C., 2020. Provenance of "after the fact" harmonised community-based demographic and HIV surveillance data from ALPHA cohorts. London School of Hygiene & Tropical Medicine.

Keilman, N., 1988. Recent trends in family and household composition in europe. Eur. J. Popul. 3, 297–325.

Kelly, C.A., Crampin, A.C., Mortimer, K., Dube, A., Malava, J., Johnston, D., Unterhalter, E., Glynn, J.R., 2018. From kitchen to classroom: Assessing the impact of cleaner burning biomass-fuelled cookstoves on primary school attendance in Karonga district, northern Malawi. PLoS One 13, e0193376.

Kenward, M.G., Carpenter, J., 2007. Multiple imputation: Current perspectives. Stat. Methods Med. Res.

Kgadima, N.P., Leburu, G.E., 2022. Attitude Toward Lobola in Remarriage Following Divorce in African Communities. Fudan J. Humanit. Soc. Sci. 16, 89–103.

Killian, M.O., Cimino, A.N., Weller, B.E., Hyun Seo, C., 2019. A Systematic Review of Latent Variable Mixture Modeling Research in Social Work Journals. J. Evid. Based. Soc. Work 16, 192–210.

Kim, H.-Y., Harling, G., Vandormael, A., Tomita, A., Cuadros, D.F., B€ Arnighausen, T., Tanser, F., 2020. HIV seroconcordance among heterosexual couples in rural KwaZulu-Natal, South Africa: a population-based analysis.

Kim, M.H., Mazenga, A.C., Yu, X., Ahmed, S., Paul, M.E., Kazembe, P.N., Abrams, E.J., 2017. High self-reported non-adherence to antiretroviral therapy amongst adolescents living with HIV in Malawi: barriers and associated factors. J. Int. AIDS Soc. 20, 21437.

Kimuna, S.R., 2005. Living arrangements and conditions of older people in Zimbabwe. Etude la Popul. Africaine 20, 143–163.

King, C., Bar-Zeev, N., Phiri, T., Beard, J., Mvula, H., Crampin, A., Heinsbroek, E., Hungerford, D., Lewycka, S., Verani, J., Whitney, C., Costello, A., Mwansambo, C., Cunliffe, N., Heyderman, R., French, N., 2020. Population impact and effectiveness of sequential 13-valent pneumococcal conjugate and monovalent rotavirus vaccine introduction on infant mortality: prospective birth cohort studies from Malawi. BMJ Glob. Heal. 5, e002669.

Kok, M.C., van Eldik, Z., Kakal, T., Munthali, A., Menon, J.A., Pires, P., Baatsen, P., van der Kwaak, A., 2021. Being dragged into adulthood? Young people's agency concerning sex, relationships and marriage in Malawi, Mozambique and Zambia. Cult. Heal. Sex. 1–26.

Korinek, K., Punpuing, S., 2012. The Effect of Household and Community on School Attrition: An Analysis of Thai Youth. Comp. Educ. Rev. 56, 474–510.

Koster, J., 2018. Family ties: The multilevel effects of households and kinship on the networks of individuals. R. Soc. Open Sci. 5.

Kreniske, P., Grilo, S., Nakyanjo, N., Nalugoda, F., Wolfe, J., Santelli, J.S., 2019. Narrating the Transition to Adulthood for Youth in Uganda: Leaving School, Mobility, Risky Occupations, and HIV. Heal. Educ. Behav. 46, 550–558.

Kriel, A., Randall, S., Coast, E., de Clercq, B., 2014. From Design to Practice: How can large-scale household surveys better represent the complexities of the social units under investigation? African Popul. Stud. 28, 1309–1323.

Kwankye, S.O., 2012. Independent North-South Child Migration as a Parental Investment in Northern Ghana. Popul. Space Place 18, 535–550.

Larmarange, J., Mossong, J., Bärnighausen, T., Newell, M.L., 2015. Participation Dynamics in Population-Based Longitudinal HIV Surveillance in Rural South Africa. PLoS One 10,

e0123345.

Lawrence Gould, A., Boye, M.E., Crowther, M.J., Ibrahim, J.G., Quartey, G., Micallef, S., Bois, F.Y., 2015. Joint modeling of survival and longitudinal non-survival data: Current methods and issues. Report of the DIA Bayesian joint modeling working group. Stat. Med. 34, 2181–2195.

Lazar, A., Jin, L., Spurlock, C.A., Wu, K., Sim, A., 2017. Data quality challenges with missing values and mixed types in joint sequence analysis. In: Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017. Institute of Electrical and Electronics Engineers Inc., pp. 2620–2627.

Lee, J., Kubik, M.Y., Fulkerson, J.A., Kohli, N., Garwick, A.E., 2020. The Identification of Family Social Environment Typologies Using Latent Class Analysis: Implications for Future Family-Focused Research. J. Fam. Nurs. 26, 26–37.

Lesclingand, M., Hertrich, V., 2017. When Girls Take the Lead: Adolescent Girls' Migration in Mali. Population (Paris). 72, 63–92.

Lesnard, L., 2006. Optimal Matching and Social Sciences.

Liao, T.F., 2004. Estimating household structure in ancient China by using historical data: a latent class analysis of partially missing patterns. J. R. Stat. Soc. Ser. A (Statistics Soc. 167, 125–139.

Liao, T.F., 2021. Using Sequence Analysis to Quantify How Strongly Life Courses Are Linked. Sociol. Sci. 8, 48–72.

Liao, T.F., Bolano, D., Brzinsky-Fay, C., Cornwell, B., Fasang, A.E., Helske, S., Piccarreta, R., Raab, M., Ritschard, G., Struffolino, E., Studer, M., 2022. Sequence analysis: Its past, present, and future. Soc. Sci. Res. 107, 102772.

Linzer, D.A., Lewis, J.B., 2011. poLCA: An R package for polytomous variable latent class analysis. J. Stat. Softw. 42, 1–29.

Locke, C., Lintelo, D.J.H., 2012. Young Zambians 'waiting' for opportunities and "working towards" living well: lifecourse and aspiration in youth transitions. J. Int. Dev. 24, 777–794.

Lovblom, L.E., Briollais, L., Tomlinson, G., Perkins, B.A., 2023. 1336-P: Predicting the Progression of Stages of Neuropathy Using Routine Eye and Kidney Tests—An Application of Novel Statistical Methods in the Diabetes Control and Complications Trial (DCCT). Diabetes 72.

Lutz, W., Cuaresma, J.C., Kebede, E., Prskawetz, A., Sanderson, W.C., Striessnig, E., 2019. Education rather than age structure brings demographic dividend. Proc. Natl. Acad. Sci. U. S. A. 116, 12798–12803.

Lynch, R., Schaffnit, S., Sear, R., Sosis, R., Shaver, J., Alam, N., Blumenfield, T., Mattison, S.M., Shenk, M., 2022. Religiosity is associated with greater size, kin density, and

geographic dispersal of women's social networks in Bangladesh. Sci. Rep. 12, 18780.

Machiyama, K., Baschieri, A., Dube, A., Crampin, A.C., Glynn, J.R., French, N., Cleland, J., 2015. An Assessment of Childbearing Preferences in Northern Malawi. Stud. Fam. Plann. 46, 161–76.

Madhavan, S., Brooks, A., 2015. Family Complexity in Rural South Africa: Examining Dynamism in Children's Living Arrangement and the Role of Kin. J. Comp. Fam. Stud.

Madhavan, S., Clark, S., Araos, M., Beguy, D., 2018. Distance or location? How the geographic distribution of kin networks shapes support given to single mothers in urban Kenya. Geogr. J. 184, 75–88.

Madhavan, S., Clark, S., Beguy, D., Kabiru, C.W., Gross, M., 2017a. Moving beyond the household: Innovations in data collection on kinship. Popul. Stud. (NY). 71, 117–132.

Madhavan, S., Mee, P., Collinson, M., 2014. Kinship in Practice: Spatial Distribution of Children's Kin Networks. J. South. Afr. Stud. 40, 401–418.

Madhavan, S., Myroniuk, T.W., Kuhn, R., Collinson, M.A., 2017b. Household structure vs. composition: Understanding gendered effects on educational progress in rural South Africa. Demogr. Res. 37, 1891–1916.

Madhavan, S., Schatz, E., Clark, S., Collinson, M., 2012. Child mobility, maternal status, and household composition in rural South Africa. Demography 49, 699–718.

Madhavan, S., Schatz, E., Gómes-Olivé, F.X., Collinson, M., 2017c. Social positioning of older persons in rural South Africa: Change or stability? J. South. Afr. Stud. 43, 1293–1307.

Madhavan, S., Townsend, N.W., Garey, A.I., 2008. "Absent breadwinners": Father-child connections and paternal support in Rural South Africa. J. South. Afr. Stud. 34, 647–663.

Magadi, M.A., 2013. Migration as a Risk Factor for HIV Infection among Youths in Sub-Saharan Africa: Evidence from the DHS. Ann. Am. Acad. Pol. Soc. Sci. 648, 136–158.

Maina, B.W., Orindi, B.O., Osindo, J., Ziraba, A.K., 2020. Depressive symptoms as predictors of sexual experiences among very young adolescent girls in slum communities in Nairobi, Kenya. Int. J. Adolesc. Youth 25, 836–848.

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

Malinga John, B., 2022. Marital Life Courses in sub-Saharan Africa: All Cause Union Dissolution, Its Timing, and Time Spent Outside Marriage. Max Planck Inst. Demogr. Res. Work. Pap.

Mandiwa, C., Namondwe, B., Munthali, M., 2021. Prevalence and correlates of comprehensive HIV/AIDS knowledge among adolescent girls and young women aged 15–24 years in Malawi: evidence from the 2015–16 Malawi demographic and health

survey. BMC Public Health 21, 1508.

Mank, I., Belesova, K., Bliefernicht, J., Traoré, I., Wilkinson, P., Danquah, I., Sauerborn, R., 2021. The Impact of Rainfall Variability on Diets and Undernutrition of Young Children in Rural Burkina Faso. Front. Public Heal. 9, 693281.

Marston, M., Nakiyingi-Miiro, J., Hosegood, V., Lutalo, T., Mtenga, B., Zaba, B., 2016. Measuring the Impact of Antiretroviral Therapy Roll-Out on Population Level Fertility in Three African Countries. PLoS One 11, e0151877.

Marston, M., Newell, M.L., Crampin, A., Lutalo, T., Musoke, R., Gregson, S., Nyamukapa, C., Nakiyingi-Miiro, J., Urassa, M., Isingo, R., Zaba, B., 2013. Is the Risk of HIV Acquisition Increased during and Immediately after Pregnancy? A Secondary Analysis of Pooled HIV Community-Based Studies from the ALPHA Network. PLoS One 8, e82219.

Maxfield, M.G., 1987. Household composition, routine activity, and victimization: A comparative analysis. J. Quant. Criminol. 3, 301–320.

McEwen, H., 2017. Nuclear power: The family in decolonial perspective and 'pro-family' politics in Africa. Dev. South. Afr. 34, 738–751.

McLean, E., Dube, A., Kalobekamo, F., Slaymaker, E., Crampin, A.C., Sear, R., 2023. Local and long-distance migration among young people in rural Malawi: importance of age, sex and family. Wellcome Open Res. 8, 211.

McLean, E., Price, A., Chihana, M., Kayuni, N., Marston, M., Koole, O., Zaba, B., Crampin, A., 2017. Changes in Fertility at the Population Level in the Era of ART in Rural Malawi. J. Acquir. Immune Defic. Syndr. 75, 391–398.

McLean, E., Price, A.J., Palla, L., Slaymaker, E., Amoah, A., Crampin, A.C., Dube, A., Kalobekamo, F., Sear, R., 2021a. 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.

McLean, E., Sironi, M., Crampin, A.C., Slaymaker, E., Dube, A., Sear, R., 2021b. Transitions to adulthood in rural Malawi in the 21st century using sequence analysis. In: International Population Conference. Hyderabad, India.

McParland, D., Gormley, I.C., McCormick, T.H., Clark, S.J., Kabudula, C.W., Collinson, M.A., 2014. Clustering South African households based on their asset status using latent variable models. Ann. Appl. Stat. 8, 747–776.

Melnikas, A.J., Mulauzi, N., Mkandawire, J., Amin, S., 2021. Perceptions of minimum age at marriage laws and their enforcement: qualitative evidence from Malawi. BMC Public Health 21, 1–12.

Melnikas, A.J., Mulauzi, N., Mkandawire, J., Saul, G., Amin, S., 2022. "Then I will say that

we have to marry each other": A qualitative view of premarital pregnancy as a driver of child marriage in Malawi. Afr. J. Reprod. Health 26, 55–63.

Mensch, B.S., Soler-Hampejsek, E., Kelly, C.A., Hewett, P.C., Grant, M.J., 2014. Challenges in Measuring the Sequencing of Life Events Among Adolescents in Malawi: A Cautionary Note. Demography 51, 277–285.

Mikolai, J., Lyons-Amos, M., 2017. Longitudinal methods for life course research: A comparison of sequence analysis, latent class growth models, and multi-state event history models for studying partnership transitions. Longit. Life Course Stud. 8, 191–208.

Mitchell, K.S., 2013. Pathways of children's long-term living arrangements: A latent class analysis. Soc. Sci. Res. 42, 1284–96.

Mkandawire, P., Tenkorang, E., Luginaah, I.N., 2013. Orphan status and time to first sex among adolescents in northern Malawi. AIDS Behav. 17, 939–50.

Molesworth, A.M., Ndhlovu, R., Banda, E., Saul, J., Ngwira, B., Glynn, J.R., Crampin, A.C., French, N., 2010. High accuracy of home-based community rapid HIV testing in rural Malawi. J. Acquir. Immune Defic. Syndr. 55, 625–30.

Molotsky, A., 2019. Income Shocks and Partnership Formation: Evidence from Malawi. Stud. Fam. Plann. 50, 219–242.

Moore, S.E., 2020. Using longitudinal data to understand nutrition and health interactions in rural Gambia. Ann. Hum. Biol. 47, 125–131.

Mukuna, R.K., 2020. Exploring Basotho teenage fathers' experiences of early fatherhood at South African rural high schools. J. Psychol. Africa 30, 348–353.

Muniina, P.N., 2016. Household survival and changes in characteristics of households in rural South-western Uganda through the period of 1989 to 2008.

Munthali, A.C., Biddlecom, A., Zulu, E.M., 2008. National Survey of Adolescents, 2004: Malawi [WWW Document]. Inter-university Consort. Polit. Soc. Res. [distributor]. URL https://www.icpsr.umich.edu/web/DSDR/studies/22410

Munthali, A.C., Zulu, E.M., 2007. The timing and role of initiation rites in preparing young people for adolescence and responsible sexual and reproductive behaviour in Malawi. Afr. J. Reprod. Health 11, 150–67.

Murtaza, S.S., Kolpak, P., Bener, A., Jha, P., 2018. Automated verbal autopsy classification: Using one-against-all ensemble method and Naïve Bayes classifier. Gates Open Res. 2.

Mutangdura, G., 2004. Women and Land Tenure Rights in Southern Africa: A Human Rights-Based Approach. In: Land in Africa: Market Asset or Secure Livelihood Conference on Gender,Land Rights and Inheritance, London. London.

Mutisya, M., Ngware, M.W., Kabiru, C.W., Kandala, N. bakwin, 2016. The effect of education

on household food security in two informal urban settlements in Kenya: a longitudinal analysis. Food Secur. 8, 743–756.

Mwambene, L., 2012. Custody Disputes Under African Customary Family Law in Malawi: Adaptability to Change? Int. J. Law Policy Fam. 26, 127–142.

Mweeba, O., Mann, G., 2020. Young marriage, parenthood and divorce in Zambia. Research Report. Oxford: Young Lives.

Myroniuk, T.W., 2018. The duration of residence spells among Malawians: the role of established family and friend connections at migrants' destinations. J. Ethn. Migr. Stud. 44, 887–907.

National Statistical Office, 2016. National Statistical Office, Zomba. Malawi Demographic and Health Survey 2015-16 Key Indicators Report.

National Statistical Office, 2019. Malawi in Figures.

Ngom, P., Magadi, M.A., Owuor, T., 2003. Parental presence and adolescent reproductive health among the Nairobi urban poor. J. Adolesc. Heal. 33, 369–377.

Nguemdjo, U., Ventelou, B., 2021. How Does Migration Affect Under-5 Mortality in Rural areas? Evidence from Niakhar, Senegal. Population (Paris). 76, 341–368.

Nguyen, D.T.N., Hughes, S., Egger, S., Lamontagne, D.S., Simms, K., Castle, P.E., Canfell, K., 2019. Risk of childhood mortality associated with death of a mother in low-and-middle-income countries: A systematic review and meta-analysis. BMC Public Health 19, 1–21.

Nguyen, N., Powers, K.A., Miller, W.C., Howard, A.G., Halpern, C.T., Hughes, J.P., Wang, J., Twine, R., Gomez-Olive, F.X., Macphail, C., Kahn, K., Pettifor, A.E., 2019. Sexual Partner Types and Incident HIV Infection among Rural South African Adolescent Girls and Young Women Enrolled in HPTN 068: A Latent Class Analysis. J. Acquir. Immune Defic. Syndr. 82, 24–33.

Ntshebe, O., Channon, A.A., Hosegood, V., 2019. Household composition and child health in Botswana. BMC Public Health 19, 1–13.

Nyirenda, M., 2014. Ageing with HIV: An investigation of the health and well-being of older people in a rural South African population with a severe HIV epidemic. University of Southampton.

Nylund-Gibson, K., Choi, A.Y., 2018. Ten frequently asked questions about latent class analysis. Transl. Issues Psychol. Sci. 4, 440–461.

Nylund, K.L., Asparouhov, T., Muthén, B.O., 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Struct. Equ. Model. 14, 535–569.

Odhiambo, F.O., Laserson, K.F., Sewe, M., Hamel, M.J., Feikin, D.R., Adazu, K., Ogwang, S., Obor, D., Amek, N., Bayoh, N., Ombok, M., Lindblade, K., Desai, M., ter Kuile, F.,

Phillips-Howard, P., van Eijk, A.M., Rosen, D., Hightower, A., Ofware, P., Muttai, H., Nahlen, B., DeCock, K., Slutsker, L., Breiman, R.F., Vulule, J.M., 2012. Profile: The KEMRI/CDC Health and Demographic Surveillance System--Western Kenya. Int. J. Epidemiol. 41, 977–987.

Odimegwu, C., Mkwananzi, S., 2016. Factors Associated with Teen Pregnancy in sub-Saharan Africa: A Multi-Country Cross-Sectional Study. Afr. J. Reprod. Health 20, 94–107.

Odimegwu, C.O., Akinyemi, J.O., De Wet, N., 2017. Premarital birth, children's sex composition and marital instability among women in sub-Saharan Africa. J. Popul. Res. 34, 327–346.

Oduro, A.R., Wak, G., Azongo, D., Debpuur, C., Wontuo, P., Kondayire, F., Welaga, P., Bawah, A., Nazzar, A., Williams, J., Hodgson, A., Binka, F., 2012. Profile of the Navrongo Health and Demographic Surveillance System. Int. J. Epidemiol. 41, 968–976.

Oduro, M.S., Iddi, S., Asiedu, L., Asiki, G., Kadengye, D.T., 2022. A multi-state transition model for child stunting in two urban slum settlements of Nairobi: a longitudinal analysis, 2011-2014 and for the Nairobi Urban Health and Demographic Surveillance System „ . medRxiv 2022.07.26.22278058.

Oketch, M., Mutisya, M., Sagwe, J., 2012. Do poverty dynamics explain the shift to an informal private schooling system in the wake of free public primary education in Nairobi slums? London Rev. Educ. 10, 3–17.

Oris, M., Ritschard, G., 2014. Sequence Analysis and Transition to Adulthood: An Exploration of the Access to Reproduction in Nineteenth-Century East Belgium. In: Life Course Research and Social Policies. Springer Science and Business Media B.V., pp. 151–167.

Palamuleni, M.E., 2011. Socioeconomic determinants of age at marriage in Malawi, International Journal of Sociology and Anthropology.

Parrot, F., Nkhata, M., Mwandosya, B., Kapira, G., Ndovi, A., Makoka, D., Gondwe, L., Crampin, A.C., 2015. Portraying fathers : reproductive journeys in Malawi 2, 154–167.

Pelletier, D., Bignami-Van Assche, S., Simard-Gendron, A., 2020. Measuring Life Course Complexity with Dynamic Sequence Analysis. Soc. Indic. Res. 152, 1127–1151.

Perez-Heydrich, C., Furgurson, J.M., Giebultowicz, S., Winston, J.J., Yunus, M., Streatfield, P.K., Emch, M., 2013. Social and spatial processes associated with childhood diarrheal disease in Matlab, Bangladesh. Heal. Place 19, 45–52.

Pérez, W., Ekholm Selling, K., Zelaya Blandón, E., Peña, R., Contreras, M., Persson, L.Å., Sysoev, O., Källestål, C., 2021. Trends and factors related to adolescent pregnancies: an incidence trend and conditional inference trees analysis of northern Nicaragua

demographic surveillance data. BMC Pregnancy Childbirth 21, 749.

Perkins, K.L., 2019. Changes in Household Composition and Children's Educational Attainment. Demography 56, 525–548.

Pesando, L.M., Barban, N., Sironi, M., Furstenberg, F.F., 2021. A Sequence-Analysis Approach to the Study of the Transition to Adulthood in Low- and Middle-Income Countries. Popul. Dev. Rev. 47, 719–747.

Piguet, E., 2018. Theories of voluntary and forced migration. In: Routledge Handbook of Environmental Migration and Displacement. pp. 17–28.

Pike, I., Mojola, S.A., Kabiru, C.W., 2018. Making Sense of Marriage: Gender and the Transition to Adulthood in Nairobi, Kenya. J. Marriage Fam. 80, 1298–1313.

Pilgrim, N.A., Ahmed, S., Gray, R.H., Sekasanvu, J., Lutalo, T., Nalugoda, F., Serwadda, D., Wawer, M.J., 2014. Family structure effects on early sexual debut among adolescent girls in Rakai, Uganda. Vulnerable Child. Youth Stud. 9, 193–205.

Pisani, E., AbouZahr, C., 2010. Sharing health data: good intentions are not enough. Bull. World Health Organ.

Pison, G., Beck, B., Ndiaye, O., Diouf, P.N., Senghor, P., Duthé, G., Fleury, L., Sokhna, C., Delaunay, V., 2018. HDSS Profile: Mlomp Health and Demographic Surveillance System (Mlomp HDSS), Senegal. Int. J. Epidemiol. 47, 1025–1033.

Pison, G., Douillot, L., Kante, A.M., Ndiaye, O., Diouf, P.N., Senghor, P., Sokhna, C., Delaunay, V., 2014. Health & Demographic Surveillance System Profile: Bandafassi Health and Demographic Surveillance System (Bandafassi HDSS), Senegal. Int. J. Epidemiol. 43, 739–748.

Pollock, G., 2007. Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. Stat. Soc. 170, 167–183.

Ponnighaus, J., Fine, P.E.M., Bliss, L., Sliney, I.J., Bradley, D.J., Rees, R.J., 1987. The Lepra Evaluation Project (LEP) an epidemiological study of leprosy in Northern Malawi. I: Methods. Lepr. Rev. 58, 359–375.

Porter, L., Hao, L., Bishai, D., Serwadda, D., Wawer, M.J., Lutalo, T., Gray, R., 2004. HIV status and union dissolution in sub-Saharan Africa: The case of Rakai, Uganda. Demography 41, 465–482.

Poulin, M., Beegle, K., Xu, H., 2021. Premarital Fertility and Marital Timing in Malawi. Stud. Fam. Plann. 52, 195–216.

Price, A.J., Glynn, J., Chihana, M., Kayuni, N., Floyd, S., Slaymaker, E., Reniers, G., Zaba, B., McLean, E., Kalobekamo, F., Koole, O., Nyirenda, M., Crampin, A.C., 2017. Sustained 10-year gain in adult life expectancy following antiretroviral therapy roll-out in rural Malawi: July 2005 to June 2014. Int. J. Epidemiol. 46, 479–491.

Rabe, M., 2008. Can the "African household" be presented meaningfully in large-scale

surveys? African Sociol. Rev. 12, 167–181.

Randall, S., Coast, E., 2015. Poverty in African Households: the Limits of Survey and Census Representations. J. Dev. Stud. 51, 162–177.

Randall, S., Coast, E., Antoine, P., Compaore, N., Dial, F.-B., Fanghanel, A., Gning, S.B., Thiombiano, B.G., Golaz, V., Wandera, S.O., 2015. UN Census “Households” and Local Interpretations in Africa Since Independence. SAGE Open 5, 215824401558935.

Randall, S., Coast, E., Leone, T., 2011. Cultural constructions of the concept of household in sample surveys. Popul. Stud. (NY).

Reniers, G., 2003. Divorce and remarriage in rural Malawi. Demogr. Res.

Reniers, G., 2008. Marital strategies for regulating exposure to HIV. Demography 45, 417–438.

Reniers, G., Wamukoya, M., Urassa, M., Nyaguara, A., Nakiyingi-Miiro, J., Lutalo, T., Hosegood, V., Gregson, S., Gómez-Olivé, X., Geubbels, E., Crampin, A.C., Wringe, A., Waswa, L., Tollman, S., Todd, J., Slaymaker, E., Serwadda, D., Price, A., Oti, S., Nyirenda, M.J., Nabukalu, D., Nyamukapa, C., Nalugoda, F., Mugurungi, O., Mtenga, B., Mills, L., Michael, D., McLean, E., McGrath, N., Martin, E., Marston, M., Maquins, S., Levira, F., Kyobutungi, C., Kwaro, D., Kasamba, I., Kanjala, C., Kahn, K., Kabudula, C., Herbst, K., Gareta, D., Eaton, J.W., Clark, S.J., Church, K., Chihana, M., Calvert, C., Beguy, D., Asiki, G., Amri, S., Abdul, R., Zaba, B., 2016. Data Resource Profile: Network for Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA Network). Int. J. Epidemiol. 45, 83–93.

Reynolds, L., 2015. Category and Kin in “Crisis”: Representations of Kinship, Care, and Vulnerability in Demographic and Ethnographic Research in KwaZulu-Natal, South Africa. Stud. Comp. Int. Dev. 50, 539–560.

Risher, K.A., Cori, A., Reniers, G., Marston, M., Calvert, C., Crampin, A., Dadirai, T., Dube, A., Gregson, S., Herbst, K., Lutalo, T., Moorhouse, L., Mtenga, B., Nabukalu, D., Newton, R., Price, A.J., Tlhajoane, M., Todd, J., Tomlin, K., Urassa, M., Vandormael, A., Fraser, C., Slaymaker, E., Eaton, J.W., 2021. Age patterns of HIV incidence in eastern and southern Africa: a modelling analysis of observational population-based cohort studies. Lancet HIV 8, e429–e439.

Ritschard, G., 2021. Measuring the Nature of Individual Sequences. Sociol. Methods Res.

Roberts, D.A., Cuadros, D., Vandormael, A., Gareta, D., Barnabas, R. V, Herbst, K., Tanser, F., Akullian, A., 2022. Predicting the Risk of Human Immunodeficiency Virus Type 1 (HIV-1) Acquisition in Rural South Africa Using Geospatial Data. Clin. Infect. Dis. 75, 1224–1231.

Robson, E., 2000. Invisible carers: young people in Zimbabwe’s home-based healthcare. Area 32, 59–69.

Romero Prieto, J., Verhulst, A., Nurul, A., Delaunay, V., Jasseh, M., Khagayi, S., Pison, G., Guillot, M., Romero Prieto, J.E., Alam, N., Reniers, G., 2022. Under-Five Mortality during the Covid-19 Outbreak: Evidence from Four Demographic Surveillance Systems in Low-Income Countries.

Sartorius, K., Sartorius, B.K., Collinson, M.A., Tollman, S.M., 2014. The dynamics of household dissolution and change in socio-economic position: A survival model in a rural South Africa. Dev. South. Afr. 31, 775–795.

Sawyer, S.M., Azzopardi, P.S., Wickremarathne, D., Patton, G.C., 2018. The age of adolescence. Lancet Child Adolesc. Heal. 2, 223–228.

Schaffnit, S.B., Urassa, M., Lawson, D.W., 2019. “Child marriage” in context: exploring local attitudes towards early marriage in rural Tanzania. Sex. Reprod. Heal. matters 27, 1571304.

Schatz, E., Madhavan, S., Collinson, M., Gómez-Olivé, F.X., Ralston, M., 2015. Dependent or Productive? A New Approach to Understanding the Social Positioning of Older South Africans Through Living Arrangements. Res. Aging 37, 581–605.

Schatz, E., Ralston, M., Madhavan, S., Collinson, M.A., Gómez-Olivé, F.X., 2018. Living Arrangements, Disability and Gender of Older Adults Among Rural South Africa. Journals Gerontol. Ser. B 73, 1112–1122.

Schuyler, A.C., Edelstein, Z.R., Mathur, S., Sekasanvu, J., Nalugoda, F., Gray, R., Wawer, M.J., Serwadda, D.M., Santelli, J.S., 2017. Mobility among youth in Rakai, Uganda: Trends, characteristics, and associations with behavioural risk factors for HIV. Glob. Public Health 12, 1033–1050.

Schwanitz, K., 2017. The transition to adulthood and pathways out of the parental home: A cross-national analysis. Adv. Life Course Res. 32, 21–34.

Scott, S., Kendall, L., Gomez, P., Howie, S.R.C., Zaman, S.M.A., Ceesay, S., D’Alessandro, U., Jasseh, M., 2017. Effect of maternal death on child survival in rural West Africa: 25 years of prospective surveillance data in The Gambia. PLoS One 12, e0172286.

Scrucca, L., Fop, B., Murphy, T.B., Raftery, A.E., 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. 8, 205–233.

Sear, R., 2021. The male breadwinner nuclear family is not the “traditional” human family, and promotion of this myth may have adverse health consequences. Philos. Trans. R. Soc. B Biol. Sci.

Sennott, C., Yeatman, S., 2012. Stability and change in fertility preferences among young women in Malawi. Int. Perspect. Sex. Reprod. Health 38, 34–42.

Shoko, M., Ibisomi, L., Levin, J., Ginsburg, C., 2018. Relationship between orphanhood status, living arrangements and sexual debut: Evidence from females in middle adolescence in Southern Africa. J. Biosoc. Sci. 50, 380–396.

Shuvai Chikovore, E., Sooryamoorthy, R., 2022. Familial Factors in Early Pregnancy Among Adolescents and Young People: An Explanatory Study of Adolescents in Cape Town, South Africa. J. Comp. Fam. Stud. 53, 256–280.

Siedner, M.J., Harling, G., Derache, A., Smit, T., Khoza, T., Gunda, R., Mngomezulu, T., Gareta, D., Majozi, N., Ehlers, E., Dreyer, J., Nxumalo, S., Dayi, N., Ording-Jesperson, G., Ngwenya, N., Wong, E., Iwuji, C., Shahmanesh, M., Seeley, J., Oliveira, T. De, Ndung'u, T., Hanekom, W., Herbst, K., 2021. Protocol: Leveraging a demographic and health surveillance system for Covid-19 Surveillance in rural KwaZulu-Natal. Wellcome Open Res. 5, 1–15.

Sifuna, P., Oyugi, M., Ogutu, B., Andagalu, B., Otieno, A., Owira, V., Otsyula, N., Oyieko, J., Cowden, J., Otieno, L., Otieno, W., 2014. Health & demographic surveillance system profile: The Kombewa health and demographic surveillance system (Kombewa HDSS). Int. J. Epidemiol. 43, 1097–104.

Siqwana-Ndulo, N., 1998. Rural African family structure in the Eastern Cape Province, South Africa. J. Comp. Fam. Stud.

Smith-Greenaway, E., Koski, A., Clark, S., 2021. Women's Marital Experiences Following Premarital Fertility in Sub-Saharan Africa. J. Marriage Fam. 83, 394–408.

Snyder, A.R., McLaughlin, D.K., Findeis, J., 2006. Household Composition and Poverty among Female-Headed Households with Children: Differences by Race and Residence. Rural Sociol. 71, 597–624.

Soler-Hampejsek, E., Kangwana, B., Austrian, K., Amin, S., Psaki, S.R., 2021. Education, Child Marriage, and Work Outcomes Among Young People in Rural Malawi. J. Adolesc. Heal. 69, S57–S64.

Somefun, O.D., 2020. Family Changes and Adolescent Development in Sub-Saharan Africa. In: Family Demography and Post-2015 Development Agenda in Africa. Springer International Publishing, pp. 243–258.

Spell, S., Anglewicz, P., Kohler, H.-P., 2012. Marriage as a Mechanism: Women's Education and Wealth in Malawi. PSC Work. Pap. Ser.

Stark, L., 2018. Poverty, Consent, and Choice in Early Marriage: Ethnographic Perspectives from Urban Tanzania. Marriage Fam. Rev. 54, 565–581.

Stark, O., Bloom, D.E., 1985. The New Economics of Labor Migration. Am. Econ. Rev. 75, 173–78.

Stiffman, M.N., Schnitzer, P.G., Adam, P., Kruse, R.L., Ewigman, B.G., 2002. Household composition and risk of fatal child maltreatment. Pediatrics 109, 615–621.

Strathman, J.G., Dueker, K.J., Davis, J.S., 1994. Effects of household structure and selected travel characteristics on trip chaining. Transportation (Amst). 21, 23–45.

Studer, M., 2013. WeightedCluster Library Manual: A practical guide to creating typologies of

trajectories in the social sciences with R. Switzerland.

Studer, M., Ritschard, G., 2016. What Matters in Differences Between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures. J. R. Stat. Soc. Ser. A Stat. Soc. 179, 481–511.

Sully, E.A., 2015. An Intimate Epidemic: HIV and Marriage in Rural Uganda. Princeton, NJ : Princeton University.

Sunny, B.S., 2018. Age-for-grade heterogeneity and primary school dropout in Karonga district, northern Malawi: Causes and consequences. London School of Hygiene and Tropical Medicine.

Sunny, B.S., DeStavola, B., Dube, A., Kondowe, S., Crampin, A.C., Glynn, J.R., 2018. Does early linear growth failure influence later school performance? A cohort study in Karonga district, northern Malawi. PLoS One 13, e0200380.

Sunny, B.S., DeStavola, B., Dube, A., Price, A., Kaonga, A.M., Kondowe, S., Crampin, A.C., Glynn, J.R., 2019. Lusting, learning and lasting in school: sexual debut, school performance and dropout among adolescents in primary schools in Karonga district, northern Malawi. J. Biosoc. Sci. 51, 720–736.

Sunny, B.S., Elze, M., Chihana, M., Gondwe, L., Crampin, A.C., Munkhondya, M., Kondowe, S., Glynn, J.R., 2017. Failing to progress or progressing to fail? Age-for-grade heterogeneity and grade repetition in primary schools in Karonga district, northern Malawi. Int. J. Educ. Dev. 52, 68–80.

Sweeting, M.J., Thompson, S.G., 2011. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. Biometrical J. 53, 750–763.

Takyi, B.K., 2001. Marital instability in an african society: Exploring the factors that influence divorce processes in ghana. Sociol. Focus 34, 77–96.

Takyi, B.K., Broughton, C.L., 2006. Marital stability in Sub-Saharan Africa: Do women's autonomy and socioeconomic situation matter? In: Journal of Family and Economic Issues. Springer, pp. 113–132.

Tangeland, T., Aas, Ø., 2011. Household composition and the importance of experience attributes of nature based tourism activity products - A Norwegian case study of outdoor recreationists. Tour. Manag. 32, 822–832.

Tanser, F., Hosegood, V., Barnighausen, T., Herbst, K., Nyirenda, M., Muhwava, W., Newell, C., Viljoen, J., Mutevedzi, T., Newell, M.-L., 2008. Cohort Profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. Int. J. Epidemiol. 37, 956–962.

Tatem, A.J., Snow, R.W., Hay, S.I., 2006. Mapping the environmental coverage of the INDEPTH demographic surveillance system network in rural Africa. Trop. Med. Int.

Heal. 11, 1318–1326.

Temin, M., Montgomery, M.R., Barker, K.M., Engebretsen, S., 2013. Girls on the Move: Adolescent Girls & Migration in the Developing World.

Templ, M., Kanjala, C., Siems, I., 2022. Privacy of Study Participants in Open-access Health and Demographic Surveillance System Data: Requirements Analysis for Data Anonymization. JMIR Public Heal. Surveill. 8, e34472.

Tenkorang, E.Y., Adjei, J.K., 2014. Household living arrangements and transition to sexual debut among young people in Ghana. Sex Educ. 15, 1–18.

Tilson, D., Larsen, U., 2000. Divorce in Ethiopia: The impact of early marriage and childlessness. J. Biosoc. Sci. 32, 355–372.

Traino, K.A., Sharkey, C.M., Perez, M.N., Bakula, D.M., Roberts, C.M., Chaney, J.M., Mullins, L.L., 2021. Health Care Utilization, Transition Readiness, and Quality of Life: A Latent Class Analysis. J. Pediatr. Psychol. 46, 197–207.

Trinitapoli, J., Yeatman, S., Fledderjohann, J., 2014. Sibling support and the educational prospects of young adults in Malawi. Demogr. Res. 30, 547–578.

Vail, L., 1989. The Creation of Tribalism in Southern Africa. Univerity of California Press, p. 153.

van Blerk, L., Ansell, N., 2006. Children's Experiences of Migration: Moving in the Wake of AIDS in Southern Africa. Environ. Plan. D Soc. Sp. 24, 449–471.

Van de Walle, E. (Ed.), 2006. African Households: Censuses and Surveys, 1st ed. Routledge, New York.

Vandormael, A., Newell, M.L., Bärnighausen, T., Tanser, F., 2014. Use of antiretroviral therapy in households and risk of HIV acquisition in rural KwaZulu-Natal, South Africa, 2004-12: A prospective cohort study. Lancet Glob. Heal. 2, 2004–2016.

Wanyua, S., Ndemwa, M., Goto, K., Tanaka, J., K'Opiyo, J., Okumu, S., Diela, P., Kaneko, S., Karama, M., Ichinose, Y., Shimada, M., 2013. Profile: The Mbita Health and Demographic Surveillance System. Int. J. Epidemiol. 42, 1678–1685.

Wasko, Z., Dambach, P., Kynast-Wolf, G., Stieglbauer, G., Zabré, P., Bagagnan, C., Schoeps, A., Souares, A., Winkler, V., 2022. Ethnic diversity and mortality in northwest Burkina Faso: An analysis of the Nouna health and demographic surveillance system from 2000 to 2012. PLOS Glob. Public Heal. 2, e0000267.

Weller, B.E., Bowen, N.K., Faubert, S.J., 2020. Latent Class Analysis: A Guide to Best Practice. J. Black Psychol. 46, 287–311.

Werner, L.K., Ludwig, J.-O., Sie, A., Bagagnan, C.H., Zabré, P., Vandormael, A., Harling, G., De Neve, J.-W., Fink, G., 2022. Health and economic benefits of secondary education in the context of poverty: Evidence from Burkina Faso. PLoS One 17, e0270246.

Whitehead, A., Hashim, I., Iverson, V., 2005. Child migration, child agency and

intergenerational relations in Africa and South Asia. Sussex, UK.

Widmer, E.D., Aeby, G., Sapin, M., 2013. Collecting family network data. Int. Rev. Sociol. 23, 27–46.

Wild, L.G., 2018. Grandparental involvement and South African adolescents' emotional and behavioural health: a summary of research findings. Contemp. Soc. Sci.

Williams, J., Schatz, E., Clark, B., Collinson, M., Clark, S., Menken, J., Kahn, K., Tollman, S., 2010. Improving public health training and research capacity in Africa: a replicable model for linking training to health and socio-demographic surveillance data. Glob. Health Action 3, 5287.

Xu, H., Mberu, B.U., Goldberg, R.E., Luke, N., 2013. Dimensions of Rural-to-Urban Migration and Premarital Pregnancy in Kenya. Ann. Am. Acad. Pol. Soc. Sci. 648, 104–119.

Yakubu, I., Salisu, W.J., 2018. Determinants of adolescent pregnancy in sub-Saharan Africa: a systematic review. Reprod. Health 15, 15.

Yamauchi, F., Buthelezi, T., Velia, M., 2008. Impacts of Prime-age Adult Mortality on Labour Supply: Evidence from Adolescents and Women in South Africa. Oxf. Bull. Econ. Stat. 70, 375–398.

Yanagisako, S.J., 1979. Family and Household: The Analysis of Domestic Groups. Annu. Rev. Anthropol. 8, 161–205.

Ye, Y., Wamukoya, M., Ezeh, A., Emina, J.B.O., Sankoh, O., 2012. Health and demographic surveillance systems: A step towards full civil registration and vital statistics system in sub-Sahara Africa? BMC Public Health 12, 741.

Yeatman, S., Chilungo, A., Lungu, S., Namadingo, H., Trinitapoli, J., 2019. Tsogolo la Thanzi : A Longitudinal Study of Young Adults Living in Malawi's HIV Epidemic. Stud. Fam. Plann. 50, 71–84.

Yotebieng, K.A., Forcone, T., 2018. The household in flux: Plasticity complicates the unit of analysis. Anthropol. Action 25, 13–22.

Young, L., Ansell, N., Lane, K., 2006. Imagining migration: placing children's understanding of "moving house" in southern Africa. Geoforum 37, 256–272.

Zietz, S., de Hoop, J., Handa, S., 2018. The role of productive activities in the lives of adolescents: Photovoice evidence from Malawi. Child. Youth Serv. Rev. 86, 246–255.

Zimmer, Z., 2009. Household Composition Among Elders in Sub-Saharan Africa in the Context of HIV/AIDS. J. Marriage Fam. 71, 1086–1099.

Ziraba, A.K., 2013. Adult mortality and its impact on children in two informal settlements in Nairobi, Kenya. London School of Hygiene & Tropical Medicine.